UNIVERSITE SULTAN MOULAY SLIMANE

Faculté des Sciences et Techniques

Béni-Mellal

Centre d'Études Doctorales : Sciences et Techniques

Formation Doctorale : Mathématiques et Physiques Appliquées

THÈSE

Présentée par

EL MASSARI HAKIM

Pour l'obtention du grade de

DOCTEUR

Discipline **:** *INFORMATIQUE*

Spécialité : *INFORMATIQUE*

Proposition d'un modèle de prédiction basé sur Machine Learning et le web sémantique

Soutenue le Mercredi 14 Juin 2023 à 10h00 devant la commission d'examen :

Pr. Mohamed BAHAJ	Professeur, Faculté des Sciences et Techniques, Université Hassan I, Settat, Maroc	Président
Pr. Abderrahim BENI-HSSANE	Professeur, Faculté des Sciences, Université Chouaib Doukkali, El Jadida, Maroc	Rapporteur
Pr. Mostafa SAADI	Professeur, ENSA, Université Sultan Moulay Slimane, Khouribga, Maroc	Rapporteur
Pr. Lahcen MOUMOUN	Professeur, ENSA, Université Hassan I, Berrechid, Maroc	Rapporteur
Pr. Mohamed AMNAI	Professeur, Faculté des Sciences, Université Ibn Tofail, Kénitra, Maroc	Examinateur
Pr. Ali OUACHA	Professeur, Faculté des Sciences, Université Mohammed V, Rabat, Maroc	Examinateur
Pr. Noreddine GHERABI	Professeur, ENSA, Université Sultan Moulay Slimane, Khouribga, Maroc	Directeur de Thèse

Dédicaces

Je dédie ce travail

A ma maman qui m'a soutenu et encouragé durant ces années d'études.

A l'homme, mon précieux offre du dieu, qui doit ma vie, ma réussite et tout mon respect mon cher père.

Zahra EL-FELLAH** et **Mohamed EL MASSARI

Qu'ils trouvent ici le témoignage de ma profonde reconnaissance.

A mes sœurs qui ont partagé avec moi tous les moments d'émotion lors de la réalisation de ce travail. Ils m'ont chaleureusement supporté et encouragé tout au long de mon parcours.

A ma famille, mes proches et à ceux qui me donnent de l'amour et de la vivacité.

A tous mes amis qui m'ont toujours encouragé, et à qui je souhaite plus de succès.

A tous ceux que j'aime.

Merci !

Remerciements

J’adresse mes sincère remerciements à mon directeur de thèse **Pr. Noreddine Gherabi** pour leur confiance et leur assistance durant toute cette période. Vous ne m’avez pas juste donné l’opportunité d’obtenir ce prestigieux diplôme, vous avez participé au développement personnel de l’homme que je suis grâce à vos conseils et leçons de vie.

Je remercie **Pr. Mohamed BAHAJ** de m’avoir fait l’honneur de présider le jury de soutenance de cette thèse.

Mes remerciements vont à l’endroit du **Pr. Abderrahim BENI-HSSANE**, **Pr. Lahcen MOUMOUN** et **Pr. Mostafa SAADI** pour avoir accepté d’être les rapporteurs de cette thèse. Vos remarques et suggestions ont été d’un grand apport à ce travail.

J’adresse mes remerciements à **Pr. Mohamed AMNAI** et **Pr. Ali OUACHA** les examinateurs de cette thèse d’avoir accepté d’examiner mon travail et de faire partie de mon jury.

Je tiens à remercier tout le personnel de l’Ecole Nationale des Sciences Appliquées de Khouribga sans exception (Directeur : **Pr. SAJIEDDINE Mohammed**, Directeur Adjoint : **Pr. KADIRI Moulay Sadik**, Secrétaire générale : **Mme JALIL Amina**, Corps professoral, Staff Administratif et Technique) pour leurs conseils, leurs recommandations, leurs soutiens et leurs encouragements tout au long de cette thèse.

Je tiens à remercier tous les membres de l’équipe doctorale pour leur amitié et leur soutien constant.

Je remercie finalement toutes les personnes qui m’ont aidé de près ou de loin, à la réalisation de ce travail.

Résumé

De nos jours, la technologie s'est améliorée dans le monde entier et est devenue une partie essentielle de notre vie. Elle aide les médecins à analyser et à diagnostiquer les problèmes médicaux et les maladies. A l'aide de l'intelligence artificielle en médecine, la science est devenue très demandée aujourd'hui. L'utilisation de l'intelligence artificielle dans de nombreux secteurs se généralise, car elle contribue à améliorer les soins de santé de plusieurs façons. Cependant, le projet d'IA est vulnérable à certains types de problèmes de santé, tels que les données non structurées, le temps de retard, etc. Par conséquent, de nouvelles approches basées sur l'ontologie doivent être intégrées, par exemple, la prédiction des maladies à l'aide de techniques d'ontologie et d'apprentissage automatique.

Les ontologies peuvent soutenir le diagnostic des maladies, en particulier en raison de leur capacité inhérente à traiter l'interopérabilité sémantique. D'un côté, l'ontologie est l'une des approches les plus adoptées pour gérer, organiser et extraire des données au cours des décennies précédentes. C'est une méthode de représentation de données qui a été mise en œuvre avec succès dans une variété de domaines, en particulier le domaine médical. Il est important en informatique en raison de sa capacité à représenter divers concepts et leurs relations dans différentes disciplines. Une approche basée sur l'ontologie pour identifier les patients atteints de maladies chroniques est une approche sémantique avec des contraintes définies, des concepts et des relations prédéfinis, y compris son propre vocabulaire. Ce modèle d'ontologie prend des requêtes, communique avec la base de connaissances et analyse les dossiers des patients par le biais d'annotations sémantiques et, dans ce cas, identifie les patients atteints de maladies chroniques ou mortelles en intégrant des langages sémantiques. L'approche basée sur l'ontologie utilise également diverses fonctions telles que la gestion terminologique, l'intégration et le partage de données, la réutilisation des connaissances et l'aide à la décision. En réalité, aucune ontologie unique n'est suffisante pour répondre aux demandes croissantes des soins de santé d'aujourd'hui, et les ontologies doivent être intégrées à des algorithmes d'apprentissage automatique pour prendre en charge l'intégration et l'analyse des données.

Dans ce contexte, les objectifs de nos travaux de recherche s'articulent autour de l'intégration de l'ontologie avec l'apprentissage automatique pour la prédiction des maladies dans le domaine de la santé. Nous avons introduit une nouvelle approche consistant à combiner l'apprentissage automatique et le Web Sémantique. D'une part, les algorithmes d'apprentissage automatique apprennent de manière autonome à effectuer une tâche ou à faire des prédictions à partir de données et améliorent leurs performances dans le temps, tandis que le Web sémantique fournit plusieurs formats d'affichage des données et des connaissances ontologiques de base. La fusion des deux domaines, nous a permis de construire un modèle basé sur une ontologie capable de prédire les maladies avec une grande précision. Nous avons appliqué l'approche sur deux différentes maladies (cardiovasculaires, cancer du sein), de plus, nous avons testé l'efficacité de cette approche dans la détection des cas du COVID-19.

Mots clés : Intelligence Artificielle, Apprentissage Automatique, Web Sémantique, Ontologie, SWRL, Healthcare, Cardiovasculaire, Cancer du sein, COVID-19.

Abstract

Nowadays, technology has improved all over the world and has become an essential part of our lives. It helps doctors analyze and diagnose medical problems and diseases. With the help of artificial intelligence in medicine, science has become in great demand today. The use of artificial intelligence in many sectors is becoming more widespread, as it helps to improve healthcare in several ways. However, the AI project is vulnerable to certain kinds of health issues, such as unstructured data, lag time, etc. Therefore, new ontology-based approaches need to be integrated, for example, disease prediction using ontology and machine learning techniques.

Ontologies can support disease diagnosis, particularly due to their inherent ability to address semantic interoperability. On the one hand, ontology is one of the most adopted approaches to manage, organize and extract data in previous decades. It is a data representation method that has been successfully implemented in a variety of fields, especially the medical field. It is important in computer science because of its ability to represent various concepts and their relationships in different disciplines. An ontology-based approach to identifying patients with chronic diseases is a semantic approach with defined constraints, predefined concepts and relationships, including its own vocabulary. This ontology model takes queries, communicates with the knowledge base and analyzes patient records through semantic annotations and, in this case, identifies patients with chronic diseases by integrating semantic languages. The ontology-based approach also uses various functions such as terminology management, data integration and sharing, knowledge reuse and decision support. In reality, no single ontology is sufficient to meet the growing demands of today's healthcare, and ontologies must be integrated with machine learning algorithms to support data integration and data analysis.

In this context, the objectives of our research work revolve around the integration of ontology with machine learning for the prediction of diseases in the field of health. We have introduced a new approach of combining machine learning and the Semantic Web. On the one hand, machine learning algorithms autonomously learn to perform a task or make predictions from data and improve their performance over time, while the Semantic Web provides multiple formats for displaying data and basic ontological knowledge. The fusion of the two domains allowed us to build an ontology-based model capable of predicting diseases with high accuracy. We applied the approach to two different diseases (cardiovascular, breast cancer), in addition, we tested the effectiveness of this approach in the detection of COVID-19 cases.

Keywords: Artificial Intelligence, Machine Learning, Semantic Web, Ontology, SWRL, Healthcare, Cardiovascular, Breast Cancer, COVID-19.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

Liste des abréviations

AB	Adaboost
ANN	Réseau de neurones artificiels
CNN	Réseaux de neurones convolutifs
DALYs	Disability-adjusted life years
DT	Decision Tree
DWT	Discrete Wavelet Transform
ESC	Société européenne de cardiologie
FNAC	Fine needle aspiration cytology
GA	Genetic algorithm
GAN	Réseaux antagonistes génératifs
IA	Intelligence artificielle
IMC	Indice de masse corporelle
IRI	Identificateur de ressource internationalisé
KNN	K-Nearest-Neighbour
LR	Logistic regression
MCC	Matthews Correlation Coefficient
MLP	Multilayer Perceptron
NB	Naive Bayes
NLP	Traitement du langage naturel
OMS	Organisation mondiale de la santé
OWL	Langage d'Ontologie Web
PSO	Particle swarm optimization
RDF	Resource Description Framework
RF	Random Forest
RNN	Réseaux de neurones récurrents
ROC	Receiver Operating Characteristics
RuleML	Rule Markup Language
SVM	Support vector machine
SWRL	Langage de règles pour le web sémantique
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WDBC	Wisconsin Breast Cancer Dataset
WTCCC	Welcome Trust Case Control Consortium
WWW	World Wide Web
XML	Extensible Markup Language

INTRODUCTION GENERALE

« Healthcare » La santé est un élément vital de toute société, car elle joue un rôle crucial dans le bien-être physique, mental et social des individus. Cela implique le diagnostic, le traitement et la prévention des maladies et des blessures, ainsi que la promotion de modes de vie sains.

Le système de santé est un réseau complexe d'organisations, de professionnels et d'établissements qui travaillent ensemble pour fournir des services de santé au public. Il comprend les hôpitaux, les cliniques, les maisons de retraite, les pharmacies et d'autres établissements qui fournissent des soins médicaux, ainsi qu'un éventail de professionnels tels que des médecins, des infirmières et d'autres membres du personnel médical formés pour diagnostiquer et traiter les maladies.

Le système de santé comprend également des agences de santé publique, qui sont chargées de promouvoir et de protéger la santé de la population. Ces agences s'efforcent de prévenir la propagation des maladies et de promouvoir des comportements sains, tels que la vaccination et l'exercice, afin de prévenir les maladies et les blessures.

Dans l'ensemble, les soins de santé sont un élément essentiel pour assurer la santé et le bien-être des individus et des communautés. Il est important de continuer à œuvrer pour l'amélioration du système de santé afin de s'assurer que chacun a accès aux soins dont il a besoin.

Il existe diverses technologies qui sont utilisées pour diagnostiquer les maladies dans les soins de santé. Certains des plus courants incluent :

- **Tests de laboratoire** : il s'agit d'analyser des échantillons de sang, d'urine ou d'autres fluides corporels dans un laboratoire pour détecter la présence d'une maladie. Cela peut inclure des tests tels que la numération globulaire, les tests de la fonction hépatique et l'analyse d'urine.
- **Radiologie** : Cela implique l'utilisation de techniques d'imagerie telles que les rayons X, les CT scans et les IRM pour visualiser l'intérieur du corps et diagnostiquer les conditions.
- **Endoscopie** : Cela consiste à insérer une petite caméra dans le corps par la bouche, le nez ou une autre ouverture pour visualiser l'intérieur du corps et diagnostiquer les conditions. Les exemples incluent la coloscopie et l'endoscopie haute.
- **Biopsie** : Cela consiste à prélever un petit échantillon de tissu du corps et à l'examiner au microscope pour diagnostiquer les conditions. Cela peut être fait par une variété de méthodes, y compris l'aspiration à l'aiguille fine, la biopsie à l'emporte-pièce et la biopsie excisionnelle.
- **Test génétique** : Il s'agit d'analyser l'ADN d'une personne pour identifier les mutations ou les variations génétiques qui peuvent être associées à une condition particulière.
- **Dossiers de santé électroniques (DSE)** : les DSE sont des enregistrements numériques des antécédents médicaux d'une personne, y compris les diagnostics, les traitements et les résultats des tests. Ceux-ci peuvent être consultés par les fournisseurs de soins de santé pour aider au diagnostic et au traitement.

- **Intelligence artificielle (IA) et apprentissage automatique** : ces technologies sont de plus en plus utilisées pour analyser les données médicales, telles que les résultats de laboratoire, les études d'imagerie et les dossiers de santé électroniques, afin d'aider au diagnostic et à la planification du traitement.
- **Web sémantique** : le web sémantique peut être utilisé pour faciliter l'échange et l'intégration des données de santé. Par exemple, il peut être utilisé pour relier les dossiers de santé électroniques à d'autres sources de données médicales, telles que les résultats de tests de laboratoire, pour aider les prestataires de soins de santé à accéder plus facilement aux antécédents médicaux d'un patient et à les analyser. Le web sémantique peut également être utilisé pour permettre l'utilisation du traitement du langage naturel pour extraire le sens de données non structurées, telles que des notes cliniques et des articles médicaux. Cela peut aider dans des tâches telles que la recherche d'informations, l'aide à la décision et la recherche clinique. Le Web sémantique peut améliorer l'efficience et l'efficacité des soins de santé en permettant une utilisation plus efficace des données et des informations.

Contexte

Les technologies d'intelligence artificielle, de plus en plus présentes dans les entreprises modernes et la vie quotidienne, sont également appliquées de manière constante aux soins de santé. L'utilisation de l'intelligence artificielle dans les soins de santé aide les prestataires de soins de santé dans de nombreux aspects des soins aux patients et des processus administratifs, en les aidant à améliorer les solutions existantes et à surmonter les défis plus rapidement. La plupart des technologies d'IA et de soins de santé ont une forte pertinence pour le domaine de la santé, mais les tactiques qu'elles soutiennent peuvent varier considérablement entre les hôpitaux et les autres organisations de soins de santé. Et tandis que certains articles sur l'intelligence artificielle dans les soins de santé suggèrent que l'utilisation de l'intelligence artificielle dans les soins de santé peut fonctionner aussi bien ou mieux que les humains lors de certaines procédures, telles que le diagnostic de maladies, il faudra un nombre important d'années avant que l'IA dans les soins de santé remplace les humains pour un large éventail de tâches médicales.

En utilisant l'intelligence artificielle dans les soins de santé, l'utilisation la plus répandue de l'apprentissage automatique traditionnel est la médecine de précision. Pouvoir prédire quelles procédures de traitement sont susceptibles de réussir avec les patients en fonction de leur constitution et du cadre de traitement est un énorme pas en avant pour de nombreux organismes de santé. La majorité des technologies d'IA dans les soins de santé qui utilisent des applications d'apprentissage automatique et de médecine de précision nécessitent des données pour l'entraînement, dont le résultat final est connu. C'est ce qu'on appelle l'apprentissage supervisé.

L'intelligence artificielle dans les soins de santé qui utilise l'apprentissage en profondeur est également utilisée pour la reconnaissance vocale sous la forme de traitement du langage naturel. Les caractéristiques des modèles d'apprentissage en profondeur ont généralement peu de sens pour les observateurs humains et, par conséquent, les résultats du modèle peuvent être difficiles à délimiter sans une interprétation appropriée.

L'apprentissage automatique est l'une des formes les plus courantes d'intelligence artificielle dans le domaine de la santé. Il s'agit d'une technique large au cœur de nombreuses approches de l'IA et des technologies de la santé et il en existe de nombreuses versions.

Pour le secteur de la santé, l'apprentissage automatique est particulièrement précieux car il peut nous aider à comprendre les quantités massives de données de santé générées chaque jour dans les dossiers de santé électroniques. L'utilisation de l'apprentissage automatique dans les soins de santé, comme les algorithmes d'apprentissage automatique, peut nous aider à trouver des modèles et des informations qu'il serait impossible de trouver manuellement, alors que l'apprentissage automatique dans les soins de santé est de plus en plus adopté, les prestataires de soins de santé ont la possibilité d'adopter une approche plus prédictive qui crée un système plus unifié avec une prestation de soins améliorée et des processus axés sur le patient.

Les cas d'utilisation les plus courants de l'apprentissage automatique dans le domaine de la santé sont l'automatisation de la facturation médicale, l'aide à la décision clinique et l'élaboration de directives de pratique clinique au sein des systèmes de santé. Il existe de nombreux exemples notables de haut niveau de concepts d'apprentissage automatique et de soins de santé appliqués en science et en médecine. Les scientifiques des données ont développé le premier algorithme d'apprentissage en profondeur dans les soins de santé pour prédire les toxicités aiguës chez les patients recevant une radiothérapie pour des cancers de la tête et du cou. Dans les flux de travail cliniques, les données générées par l'apprentissage en profondeur dans les soins de santé peuvent identifier automatiquement des modèles complexes et offrir à un fournisseur de soins primaires une aide à la décision clinique au point de service dans le dossier de santé électronique.

Dans le domaine de la santé, les algorithmes d'apprentissage automatique peuvent être utilisés pour prédire les résultats des patients, identifier les épidémies potentielles de maladies infectieuses et aider au diagnostic et au traitement des maladies. L'apprentissage automatique peut également être utilisé pour optimiser la découverte de médicaments et améliorer l'efficacité de la tenue des dossiers médicaux. Certaines applications spécifiques de l'apprentissage automatique dans les soins de santé comprennent :

- **Modélisation prédictive** : L'apprentissage automatique peut être utilisé pour prédire la probabilité qu'un patient développe une certaine condition ou répond à un traitement particulier. Cela peut aider les fournisseurs de soins de santé à prendre des décisions plus efficaces concernant les soins aux patients.
- **Aide à la décision clinique** : les algorithmes d'apprentissage automatique peuvent être utilisés pour analyser les antécédents médicaux d'un patient et fournir des recommandations sur les options de traitement. Cela peut aider les médecins à prendre des décisions plus adaptées aux soins d'un patient et à réduire le risque d'erreurs.
- **Diagnostic de la maladie** : les algorithmes d'apprentissage automatique peuvent être formés pour identifier des modèles dans les images médicales, telles que les rayons X ou les IRM, qui peuvent indiquer la présence d'une maladie particulière. Cela peut aider les médecins à établir des diagnostics plus précis et à fournir un traitement plus ciblé.
- **Découverte de médicaments** : l'apprentissage automatique peut être utilisé pour analyser des composés chimiques et prédire leur efficacité potentielle en tant que

médicaments. Cela peut aider à accélérer le processus de découverte et de développement de médicaments.

- **Gestion de la santé de la population** : les algorithmes d'apprentissage automatique peuvent être utilisés pour analyser de grandes quantités de données provenant de plusieurs sources, telles que les dossiers de santé électroniques et les appareils portables, afin d'identifier les modèles et les tendances de la santé de la population. Cela peut aider les fournisseurs de soins de santé à concevoir des interventions et des programmes de soins préventifs plus efficaces.

L'utilisation du web sémantique et de ces technologies dans nos travaux de recherche référer à plusieurs raisons notamment la prédiction des maladies, soutenir le diagnostic des maladies… Nous donnant plus de détail sur l'utilisation des technologies du web sémantique dans les parties suivantes.

Le Web sémantique est un ensemble de normes et de technologies qui visent à rendre le World Wide Web plus structuré et lisible par machine. Dans le contexte de la santé, le Web sémantique peut être utilisé pour représenter et partager des données et des connaissances liées à la santé d'une manière plus facilement compréhensible et traitée par les ordinateurs. Cela peut permettre un large éventail d'applications, telles que les dossiers de santé électroniques, l'aide à la décision clinique et la gestion de la santé de la population.

Un aspect clé du Web sémantique est l'utilisation de vocabulaires et d'ontologies normalisés, qui fournissent un langage commun pour représenter et organiser les données et les connaissances liées aux soins de santé. Ces vocabulaires et ontologies peuvent aider à garantir que différents systèmes et ensembles de données peuvent être interopérables, ce qui signifie qu'ils peuvent être facilement partagés et intégrés les uns aux autres.

Un autre aspect important du Web sémantique dans le domaine de la santé est l'utilisation de données liées, qui fait référence à la pratique consistant à lier des données et des connaissances connexes à l'aide d'adresses Web normalisées (URI). Cela peut aider à créer un réseau d'informations liées aux soins de santé plus interconnecté, plus facilement navigable et détectable par les humains et les ordinateurs.

Dans l'ensemble, le Web sémantique a le potentiel de transformer la façon dont les données et les connaissances sur les soins de santé sont gérées et utilisées, permettant une prestation de soins de santé, une recherche et une élaboration des politiques plus efficientes et efficaces.

Dans le contexte des soins de santé, une ontologie est un vocabulaire normalisé et un ensemble de concepts utilisés pour représenter et organiser les connaissances sur un domaine particulier. Dans le cas des soins de santé, cela peut inclure des concepts tels que les maladies, les traitements et les tests de diagnostic, ainsi que les relations entre ces concepts.

Les ontologies peuvent être utilisées pour représenter les connaissances liées aux soins de santé d'une manière qui est plus facilement comprise et traitée par les ordinateurs. Cela peut permettre un large éventail d'applications, telles que les dossiers de santé électroniques, l'aide à la décision clinique et la gestion de la santé de la population.

L'un des principaux avantages de l'utilisation des ontologies dans le domaine de la santé est qu'elles peuvent aider à garantir que différents systèmes et ensembles de données sont interopérables, ce qui signifie qu'ils peuvent être facilement partagés et intégrés les uns aux

autres. Cela peut aider à réduire les obstacles au partage de données et faciliter l'échange d'informations entre les différents organismes de soins de santé et les parties prenantes.

Les ontologies peuvent également être utilisées pour soutenir la représentation et le raisonnement sur des scénarios cliniques complexes, en fournissant une représentation structurée des connaissances pertinentes et des relations entre différents concepts. Cela peut permettre le développement de systèmes d'aide à la décision clinique plus sophistiqués et d'autres applications qui reposent sur une compréhension détaillée des connaissances sous-jacentes.

Dans l'ensemble, l'utilisation d'ontologies dans les soins de santé a le potentiel d'améliorer la précision et la rapidité de la prise de décision en matière de soins de santé, ainsi que de soutenir le développement d'une prestation de soins de santé et d'une recherche plus efficace et efficiente.

Problématique et Objectifs

De nos jours, la technologie s'est améliorée dans le monde entier et est devenue une partie essentielle de notre vie. Il aide les médecins à analyser et à diagnostiquer les problèmes médicaux et les maladies. Avec l'aide de l'intelligence artificielle en médecine, la science est devenue très demandée aujourd'hui. L'utilisation de l'intelligence artificielle dans de nombreux secteurs se généralise, car elle contribue à améliorer les soins de santé de plusieurs façons. Cependant, le projet d'IA est vulnérable à certains types de problèmes de santé, tels que les données non structurées, le temps de retard, etc. Par conséquent, de nouvelles approches basées sur l'ontologie doivent être intégrées, par exemple, la prédiction des maladies à l'aide de techniques d'ontologie et d'apprentissage automatique.

Les ontologies peuvent soutenir le diagnostic des maladies, en particulier en raison de leur capacité inhérente à traiter l'interopérabilité sémantique. Par conséquent, l'approche ontologique fournit un cadre sémantique dans lequel les données des patients pour l'évaluation et la gestion de leurs maladies peuvent être saisies, et les risques, profils et recommandations dérivés. L'approche basée sur l'ontologie peut également soutenir l'évaluation de la qualité des données cliniques utilisées. Une approche basée sur l'ontologie pour identifier les patients atteints de maladies chroniques est une approche sémantique avec des contraintes définies, des concepts et des relations prédéfinis, y compris son propre vocabulaire. Ce modèle d'ontologie prend des requêtes, communique avec la base de connaissances et analyse les dossiers des patients par le biais d'annotations sémantiques et, dans ce cas, identifie les patients atteints de maladies chroniques en intégrant des langages sémantiques. L'approche basée sur l'ontologie utilise également diverses fonctions telles que la gestion terminologique, l'intégration et le partage de données, la réutilisation des connaissances et l'aide à la décision. Les ontologies doivent représenter la réalité tout en ayant une base théorique solide.

D'un autre côté, l'ontologie est l'une des approches les plus adoptées pour gérer, organiser et extraire des données au cours des décennies précédentes. C'est une méthode de représentation de données qui a été mise en œuvre avec succès dans une variété de domaines, en particulier le domaine médical. Il est important en informatique en raison de sa capacité à représenter divers concepts et leurs relations dans différentes disciplines. En réalité, aucune ontologie unique n'est suffisante pour répondre aux demandes croissantes des soins de santé

d'aujourd'hui, et les ontologies doivent être intégrées à des algorithmes d'apprentissage automatique pour prendre en charge l'intégration et l'analyse des données.

Dans ce contexte, les objectifs de nos travaux de recherche s'articulent autour de l'intégration de l'ontologie avec l'apprentissage automatique pour la prédiction des maladies dans le domaine de la santé. Nous avons introduit une nouvelle approche consistant à combiner l'apprentissage automatique et le Web Sémantique. D'une part, les algorithmes d'apprentissage automatique apprennent de manière autonome à effectuer une tâche ou à faire des prédictions à partir de données et améliorent leurs performances dans le temps, tandis que le Web sémantique fournit plusieurs formats d'affichage des données et des connaissances ontologiques de base. La fusion des deux domaines, nous a permis de construire un modèle basé sur une ontologie capable de prédire les maladies avec une grande précision. Nous avons appliqué l'approche sur deux différentes maladies (cardiovasculaires, cancer du sein), plus que ça nous avons aussi testé l'efficacité de cette approche dans la détection des cas du COVID-19.

Organisation du manuscrit

Ce mémoire de thèse est constitué de cinq chapitres :

- **Chapitre I : Définitions et concepts de base**

Ce premier chapitre introduit les concepts et les définitions de base qui seront nécessaires pour présenter l'état de l'art et les approches proposées dans les chapitres suivants.

- **Chapitre II : Revue de l'état de l'art**

Ce chapitre présente un état de l'art exhaustif sur l'emploi des ontologies et des techniques d'apprentissage automatique pour la prédiction des maladies cardiovasculaires, le cancer du sein, et la détection des cas du COVID-19. Également présente l'impact d'appliquer l'apprentissage automatique basé sur des ontologies pour avoir de meilleurs résultats.

- **Chapitre III : L'impact de l'ontologie sur la prédiction des maladies cardiovasculaires par rapport aux algorithmes d'apprentissage automatique**

Dans ce chapitre, nous avons introduit une nouvelle approche consistant à fusionner l'apprentissage automatique et le Web Sémantique. D'une part, les algorithmes d'apprentissage automatique apprennent de manière autonome à effectuer une tâche ou à faire des prédictions à partir de données et améliorent leurs performances dans le temps, tandis que le Web sémantique fournit plusieurs formats d'affichage des données et des connaissances ontologiques de base. La fusion des deux domaines, nous a permis de construire un modèle basé sur une ontologie capable de prédire les maladies cardiovasculaires avec une grande précision.

- **Chapitre VI : Intégration de l'ontologie avec l'apprentissage automatique pour prédire la présence de covid-19 en fonction des symptômes**

Dans ce chapitre, nous avons démontré comment les ontologies peuvent aider à prédire la présence de COVID-19 sur la base des symptômes, en intégrant l'ontologie et l'apprentissage automatique, en implémentant les règles de l'algorithme de l'arbre de décision dans le raisonneur d'ontologie. Une analyse comparative a été menée en évaluant les performances

du modèle en validation croisée 10 fois et en fractionnement de test en pourcentage via le logiciel d'apprentissage automatique WEKA. Les résultats sont évalués à l'aide de mesures de performance générées à partir de la matrice de confusion. Cette étude peut servir de système d'aide à la décision pour les médecins, en utilisant le modèle développé comme aide pour détecter la présence de COVID-19 chez une personne en fonction des symptômes déclarés.

- **Chapitre V : Modèle ontologique basé sur l'apprentissage automatique pour prédire le cancer du sein**

Dans ce chapitre, nous avons comparé sept approches de classification nommées ; machine à vecteurs de support (SVM), K-plus proches voisins, forêts aléatoires, réseaux de neurones artificiels (ANN) et régression logistique, en plus du modèle ontologique basé sur l'algorithme d'arbre de décision. Le jeu de données sur le cancer du sein du (Breast Cancer Wisconsin (Diagnostic) Data Set) est obtenu à partir d'une importante base de données d'apprentissage automatique appelée base de données d'apprentissage automatique UCI. La performance de l'étude est mesurée en termes d'exactitude, de précision, de rappel et de F-mesure.

Nous clôturons ce manuscrit par une synthèse de nos travaux tout en donnant une vue sur les perspectives de recherche envisagées et les nouveaux axes de recherche qui nous paraissent les plus pertinents.

CHAPITRE I - DEFINITIONS ET CONCEPTS DE BASE

I.1. Introduction

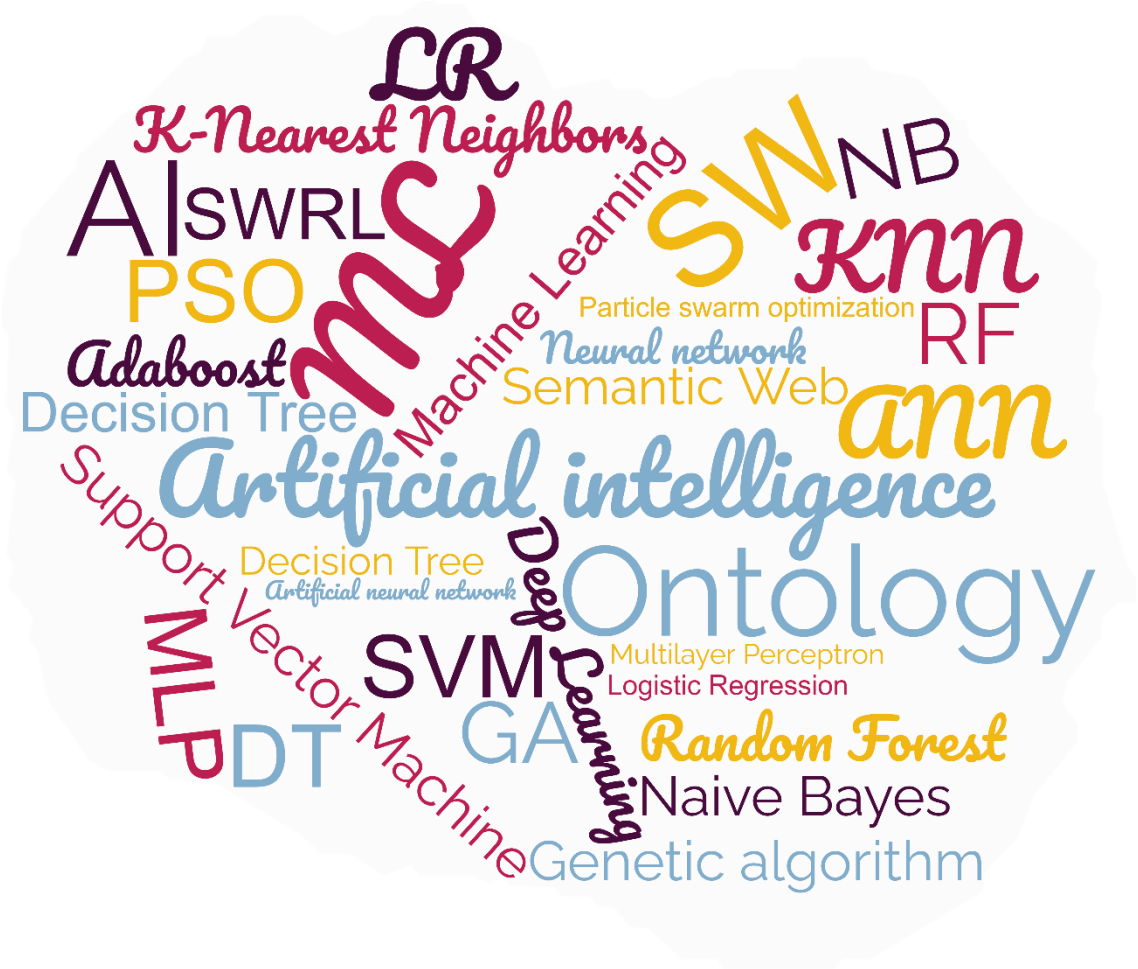

Figure I.1 - Nuage de mots des concepts et technologies utilisés dans cette thèse.

L'intelligence artificielle, l'apprentissage automatique et la science des données sont tous liés les uns aux autres. Sans surprise, ils sont souvent utilisés de manière interchangeable et confondus les uns avec les autres dans les médias populaires et la communication commerciale. Cependant, ces trois domaines sont distincts selon le contexte. La Figure I.2 montre la relation entre l'intelligence artificielle, l'apprentissage automatique et la science des données.

L'intelligence artificielle consiste à donner aux machines la capacité d'imiter le comportement humain, en particulier les fonctions cognitives. Exemples : reconnaissance faciale, conduite automatisée, tri du courrier en fonction du code postal. Il existe toute une gamme de techniques qui relèvent de l'intelligence artificielle : linguistique, traitement du langage naturel, science de la décision, biais, vision, robotique, planification, etc. L'apprentissage est une partie importante de la capacité humaine. En fait, de nombreux autres organismes vivants peuvent apprendre.

L'apprentissage automatique peut être considéré comme un sous-domaine ou l'un des outils de l'intelligence artificielle, il fournit aux machines la capacité d'apprendre de l'expérience.

L'expérience des machines se présente sous la forme de données. Les données utilisées pour enseigner aux machines sont appelées données de formation. L'apprentissage automatique bouleverse le modèle de programmation traditionnel (Figure I.3). Un programme, un ensemble d'instructions pour un ordinateur, transforme les signaux d'entrée en signaux de sortie en utilisant des règles et des relations prédéterminées. Les algorithmes d'apprentissage automatique, également appelés "apprenants", prennent à la fois l'entrée et la sortie connues (données d'apprentissage) pour déterminer un modèle pour le programme qui convertit l'entrée en sortie. Par exemple, de nombreuses organisations telles que les plateformes de médias sociaux, les sites d'évaluation ou les forums sont tenues de modérer les publications et de supprimer le contenu abusif. Comment apprendre aux machines à automatiser la suppression des contenus abusifs ? Les machines doivent recevoir des exemples de messages abusifs et non abusifs avec une indication claire de celui qui est abusif. Les apprenants généraliseront un modèle basé sur certains mots ou séquences de mots afin de conclure si le message global est abusif ou non. Le modèle peut prendre la forme d'un ensemble de règles "si - alors". Une fois que les règles ou le modèle de science des données sont développés, les machines peuvent commencer à catégoriser la disposition de tout nouveau message.

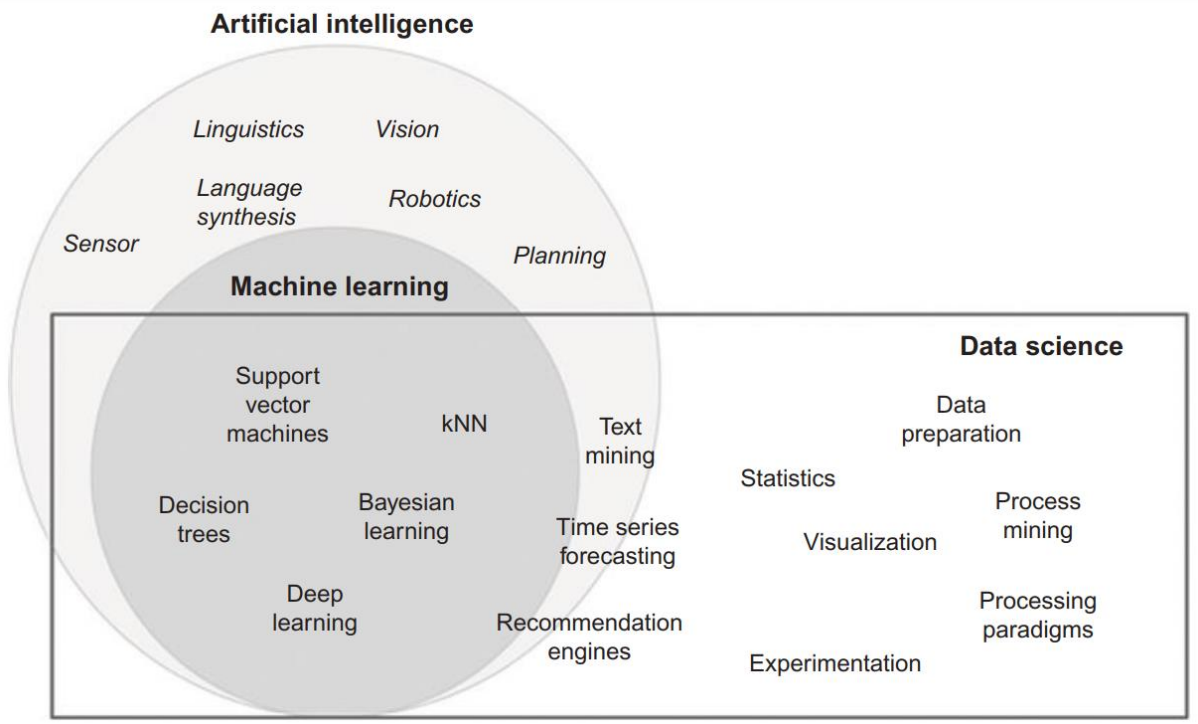

Figure I.2 - Intelligence artificielle, apprentissage automatique et science des données [1].

La science des données est l'application commerciale de l'apprentissage automatique, de l'intelligence artificielle et d'autres domaines quantitatifs tels que les statistiques, la visualisation et les mathématiques. C'est un domaine interdisciplinaire qui extrait la valeur des données. Dans le contexte de la façon dont la science des données est utilisée aujourd'hui, elle s'appuie fortement sur l'apprentissage automatique et est parfois appelée exploration de données. Des exemples de cas d'utilisation de la science des données sont : les moteurs de recommandation qui peuvent recommander des films pour un utilisateur particulier, un modèle d'alerte à la fraude qui détecte les transactions frauduleuses par carte de crédit,

trouver les clients qui vont très probablement se désabonner le mois prochain ou prédire les revenus pour le prochain trimestre.

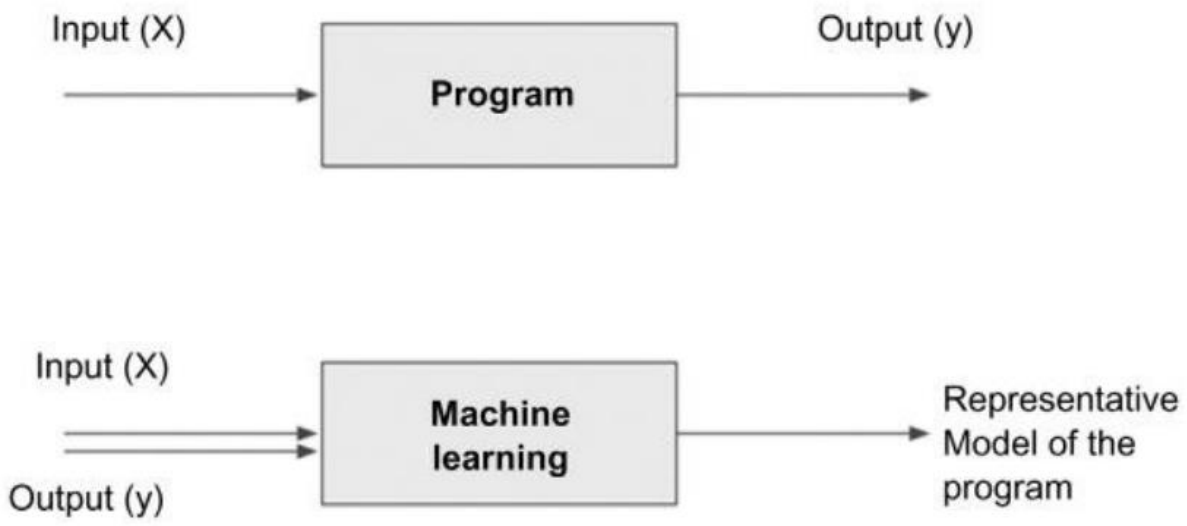

Figure I.3 - Programme traditionnel et apprentissage automatique.

Le web sémantique apporte une solution à la machine pour traiter les données en ligne. Le Web sémantique [2] est une extension du *World Wide Web (WWW)* actuel qui offre des applications programmables avec des métadonnées interprétables par machine des données en ligne. Le Web sémantique ajoute des descripteurs de données supplémentaires au contenu disponible sur le Web. Cela permet aux machines de faire des interprétations significatives de la même manière que les gens analysent les données pour prendre des décisions utiles.

L'ontologie est une description formelle des connaissances comme un ensemble de classes au sein d'un domaine spécifique et des relations qui existent entre elles [3]. Récemment, les ontologies ont attiré une attention considérable dans la conception de la connaissance du domaine des cours, de l'apprentissage en ligne, des actualités, du génie logiciel, etc.

Nous introduisons dans ce chapitre les concepts et les définitions de base qui seront nécessaires pour présenter l'état de l'art et les approches proposées dans les chapitres suivants.

I.2. Intelligence artificielle

L'intelligence artificielle, ou IA, est la simulation de l'intelligence humaine dans des machines programmées pour penser et agir comme des humains. Ces machines sont conçues pour apprendre et s'adapter à de nouvelles informations, résoudre des problèmes et prendre des décisions en fonction des données qui leur ont été fournies. L'intelligence artificielle a le potentiel de révolutionner de nombreuses industries et a un large éventail d'applications, des voitures autonomes et du diagnostic médical à la reconnaissance d'images et au traitement du langage naturel. Cependant, le développement de l'IA soulève également des préoccupations éthiques, telles que la perte potentielle d'emplois et la possibilité que l'IA soit utilisée à des fins malveillantes.

Il existe de nombreux composants différents qui peuvent être utilisés dans les systèmes d'intelligence artificielle (Figure I.4).

Certains des composants clés incluent :

1. **Algorithmes d'apprentissage automatique** : ce sont des algorithmes qui permettent à un système d'apprendre à partir de données et d'améliorer ses performances au fil du temps.
2. **Traitement du langage naturel (NLP)** : Il s'agit d'un domaine de l'IA qui vise à permettre à un système de comprendre et de générer un langage humain.
3. **Vision par ordinateur** : Il s'agit d'un domaine de l'IA qui vise à permettre à un système d'interpréter des données visuelles, telles que des images et des vidéos.
4. **Robotique** : Il s'agit d'un domaine de l'IA qui se concentre sur la conception et le contrôle de robots, qui sont des machines qui peuvent être programmées pour effectuer une variété de tâches.
5. **Représentation des connaissances et raisonnement** : il s'agit d'un domaine de l'IA qui se concentre sur la manière de représenter les connaissances et de les utiliser pour prendre des décisions et résoudre des problèmes.
6. **Prise de décision** : il s'agit d'un domaine de l'IA qui se concentre sur la conception d'algorithmes et de systèmes capables de prendre des décisions en fonction des informations disponibles.

En général, les différents composants de l'intelligence artificielle fonctionnent ensemble pour permettre à un système d'effectuer des tâches intelligentes, telles que la reconnaissance de visages sur des photos, la traduction de langues, etc.

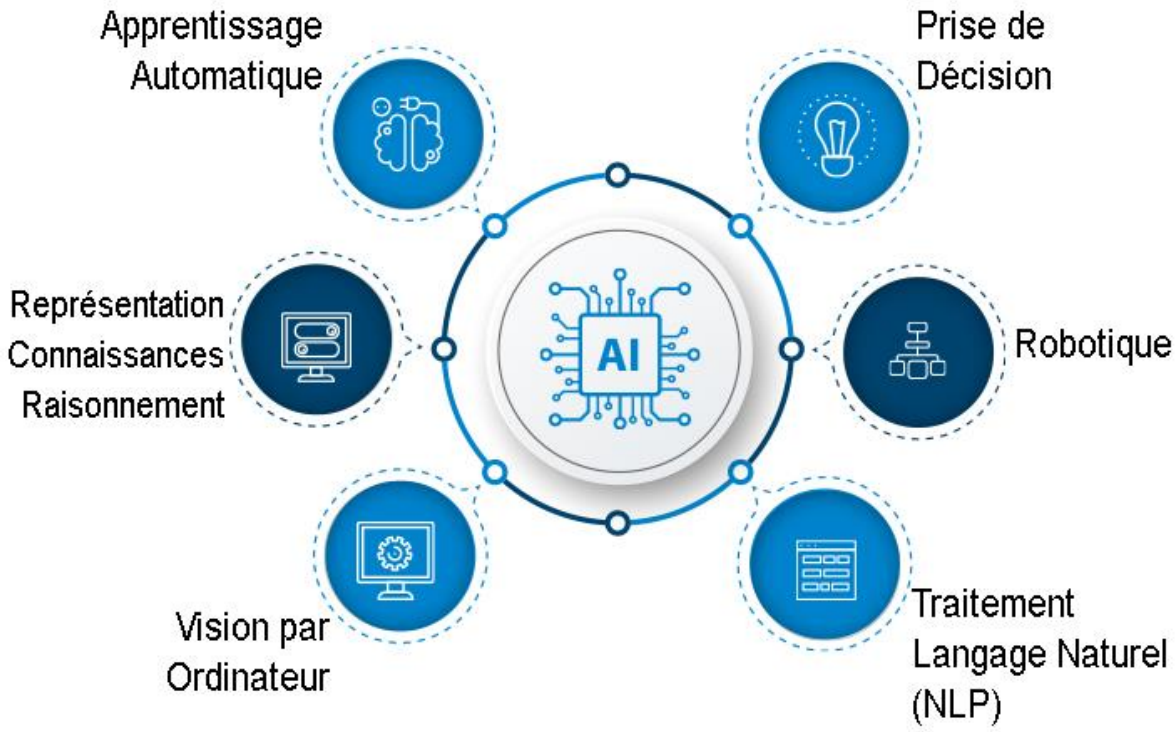

Figure I.4 - Différents composants de l'intelligence artificielle.

I.3. Apprentissage automatique

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle qui vise à permettre aux ordinateurs d'apprendre à partir de données, sans être explicitement programmés. L'objectif de l'apprentissage automatique est de créer des algorithmes capables de recevoir des données d'entrée et de les utiliser pour faire des prédictions ou prendre des mesures afin d'atteindre un objectif spécifique. Les algorithmes d'apprentissage automatique peuvent être supervisés, ce qui signifie qu'ils sont formés à l'aide de données étiquetées, ou non supervisés, ce qui signifie qu'ils sont formés à l'aide de données non structurées. Il existe de nombreux types d'algorithmes d'apprentissage automatique, notamment des arbres de décision, des forêts aléatoires, des réseaux de neurones, etc.

L'apprentissage automatique implique généralement l'utilisation d'algorithmes et de modèles statistiques pour permettre à un système d'améliorer ses performances sur une tâche spécifique au fil du temps. Cela peut impliquer l'utilisation de divers composants, tels que le prétraitement des données, l'ingénierie des fonctionnalités, l'entraînement de modèles et l'évaluation. Le prétraitement des données implique le nettoyage et le formatage des données de manière à les rendre utilisables dans un modèle d'apprentissage automatique. L'ingénierie des fonctionnalités implique la sélection et la création de fonctionnalités pertinentes à partir des données brutes qui peuvent être utilisées pour entraîner le modèle. L'entraînement de modèle implique l'utilisation de ces fonctionnalités pour entraîner le modèle, généralement à l'aide d'une variété d'algorithmes et d'hyperparamètres. Enfin, l'évaluation du modèle implique d'évaluer les performances du modèle formé sur un ensemble de données distinct pour s'assurer qu'il fait des prédictions précises.

I.3.1. Composants communs de l'apprentissage automatique

L'apprentissage automatique est une méthode permettant aux ordinateurs d'apprendre à partir de données, sans être explicitement programmés. Cela implique l'utilisation d'algorithmes et de modèles statistiques pour permettre à un système d'améliorer ses performances sur une tâche spécifique au fil du temps. Certains composants courants d'un système d'apprentissage automatique incluent : L'apprentissage automatique est une méthode permettant aux ordinateurs d'apprendre à partir de données, sans être explicitement programmés. Cela implique l'utilisation d'algorithmes et de modèles statistiques pour permettre à un système d'améliorer ses performances sur une tâche spécifique au fil du temps. Certains composants communs d'un système d'apprentissage automatique incluent :

- **Données** : les algorithmes d'apprentissage automatique nécessitent une grande quantité de données pour apprendre. Ces données sont utilisées pour entraîner le modèle et faire des prédictions sur de nouvelles données inédites.
- **Modèle** : le modèle est le composant central d'un système d'apprentissage automatique. Il s'agit d'une représentation mathématique des relations et des modèles trouvés dans les données. Le modèle est créé par l'algorithme d'apprentissage et est utilisé pour faire des prédictions sur de nouvelles données.
- **Algorithme** : un algorithme d'apprentissage est un ensemble d'instructions qui indiquent au modèle comment trouver des modèles dans les données et comment utiliser ces modèles pour faire des prédictions. Il existe de nombreux types d'algorithmes différents, chacun avec ses propres forces et faiblesses.

- **Évaluation** : Une fois que le modèle a été formé et fait des prédictions, il est important d'évaluer ses performances pour s'assurer qu'il est précis et fiable. Cela se fait généralement à l'aide d'un ensemble de données de test que le modèle n'a jamais vu auparavant.

I.3.2. Déploiement d'un modèle d'apprentissage automatique

La construction d'un modèle d'apprentissage automatique est un processus itératif. Pour un déploiement réussi, la plupart des étapes sont répétées plusieurs fois pour obtenir des résultats optimaux. Le modèle doit être maintenu après le déploiement et adapté à l'évolution de l'environnement. La Figure I.5 illustre les étapes du cycle de vie d'un modèle d'apprentissage automatique.

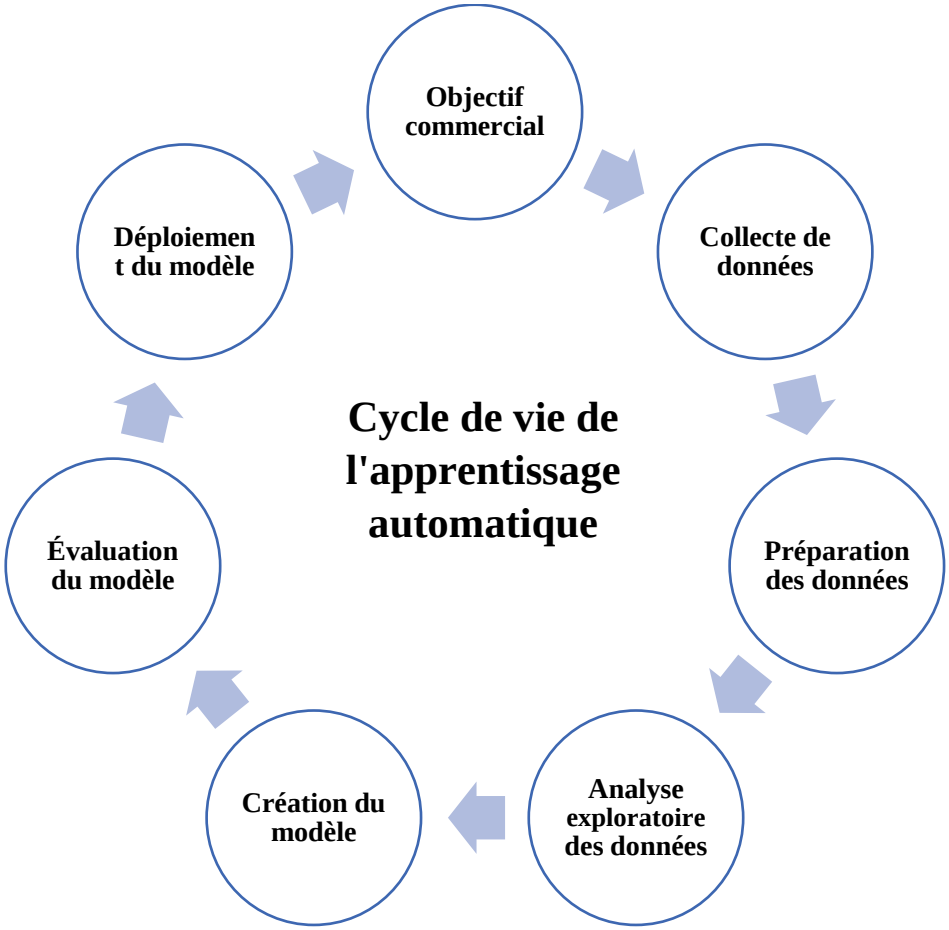

Figure I.5 - Cycle de vie de l'apprentissage automatique.

La figure ci-dessus représente toutes les étapes impliquées dans le cycle de vie de l'apprentissage automatique requis pour le développement et la mise en œuvre des modèles d'apprentissage automatique. C'est un processus cyclique parce que le processus est continu et répétitif, suivez séquentiellement une étape après une autre étape.

- ✓ **Objectif commercial / Énoncé du problème**

La première étape de chaque projet d'apprentissage automatique consiste à connaître et à comprendre l'objectif commercial. Si les exigences de l'entreprise ne sont pas claires, il sera

inutile d'aller plus loin. Comprendre les besoins de l'entreprise et connaître les clients (utilisateurs finaux) est essentiel pour créer un modèle précis.

- ✓ **Collecte / Acquisition de données**

Les données sont comme du carburant pour un algorithme d'apprentissage automatique. Ainsi, il peut être collecté à partir de différentes sources, mais nous devons nous assurer que nous collectons des données correctes et pertinentes pour les besoins de l'entreprise. Les sources de données peuvent être des sites Web de commerce électronique, des médias sociaux, des bases de données (MySQL, Oracle, DB2 et bien d'autres), etc. Il est très important d'identifier les variables indépendantes (X ou variables d'entrée) et les variables dépendantes (Y ou variable de sortie) basé sur les besoins de l'entreprise et pour collecter des données en conséquence.

Voici les étapes impliquées dans la collecte de données :

1. Identifier différentes sources.
2. Extraire des données de différentes sources.
3. Intégrez des données provenant de différentes sources pour entraîner un ensemble de données requis pour entraîner un modèle.

- ✓ **Préparation des données**

Une fois les données collectées, il est essentiel de nettoyer et de pré-traiter les données pour obtenir des résultats précis à partir de notre modèle. La conversion des données brutes en un format utilisable est la base de cette étape. Ça implique :

1. Traitement des valeurs manquantes,
2. Vérification du type de données incorrect ou des données invalides,
3. Examen des valeurs nulles,
4. Élimination des enregistrements en double.

- ✓ **Analyse exploratoire des données (EDA)**

Dans cette étape, nous explorons les données plus en profondeur pour mieux comprendre les données afin que des informations puissent être générées afin que les questions que nous avons à l'esprit puissent être traitées. Nous comprenons comment les données sont distribuées, visualisons les données, identifions les valeurs aberrantes le cas échéant, traitons ces valeurs aberrantes avec différentes techniques et analysons les modèles dans les données. Différents tracés comme l'histogramme, la boîte à moustaches, le nuage de points aident à visualiser les données.

Voici les étapes requises dans EDA :

1. Visualisation
2. Analyses statistiques
3. Traitement des valeurs aberrantes
4. Transformation des données

5. Sélection de fonctionnalité
6. Création de variables factices
7. Partitionnement des données
8. Équilibrer les ensembles de données déséquilibrés

- ✓ **Création du modèle**

À l'aide de divers algorithmes d'apprentissage automatique (régression ou classification), un ensemble de données d'entraînement préparé dans les étapes ci-dessus est utilisé pour entraîner le modèle. Ceci est complet afin que le modèle puisse comprendre différents modèles et relations dans les ensembles de données.

- ✓ **Évaluation du modèle**

Une fois que le modèle est entraîné, il est très important d'analyser les performances et l'exactitude du modèle. L'objectif principal de tout projet d'apprentissage automatique est d'atteindre une exactitude maximale du modèle. Si le modèle ne parvient pas à classer ou à prédire correctement la sortie, ce modèle particulier n'est pas considéré comme le meilleur modèle. Ainsi, l'ensemble de données de test est utilisé pour analyser les performances et l'exactitude du modèle.

Il existe cependant divers critères de mesures d'évaluation (RMSE, score d'exactitude, précision, rappel, score F1, courbe ROC, etc.) sur la base desquels le meilleur modèle est sélectionné. De plus, les problèmes de surajustement et de sous-ajustement sont résolus afin que le modèle fonctionne bien avec de nouvelles données.

- ✓ **Déploiement du modèle**

Sur la base des performances des modèles, le meilleur modèle est choisi. La fonctionnalité et la sortie des modèles doivent être accessibles aux utilisateurs finaux. Ainsi, en fonction des exigences et des objectifs commerciaux, la phase de déploiement doit être simple ou complexe. Mais nous devons nous assurer qu'il répond à l'objectif commercial en termes de précision et de vitesse minimales.

Les étapes mentionnées ci-dessus sont suivies séquentiellement lors du développement du modèle d'apprentissage automatique.

I.3.3. Différents types de paradigmes d'apprentissage automatique

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle (IA) qui se concentre sur le développement d'algorithmes et de modèles qui peuvent être formés pour apprendre automatiquement à partir de données sans être explicitement programmés. Les modèles d'apprentissage automatique sont formés sur de grands ensembles de données et utilisent ces données pour faire des prédictions ou prendre des mesures en fonction de nouvelles entrées. Ces modèles peuvent être utilisés pour un large éventail d'applications, notamment la reconnaissance d'images et de la parole, le traitement du langage naturel et l'analyse prédictive.

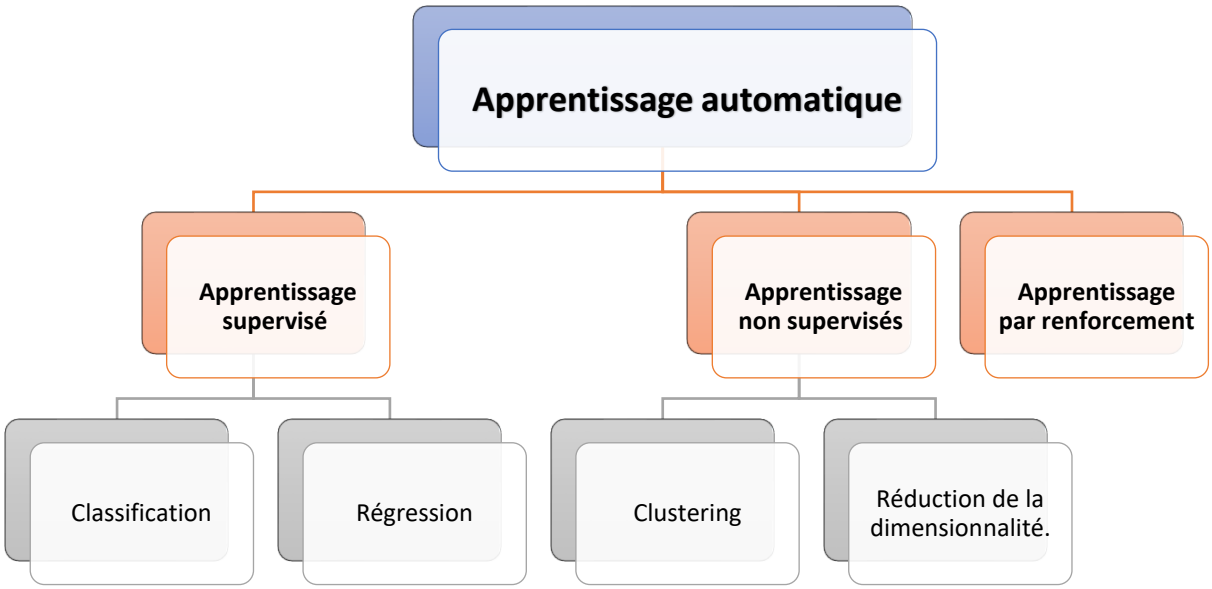

Figure I.6 - Types d'algorithmes d'apprentissage automatique.

Il existe plusieurs types de modèles d'apprentissage automatique, chacun étant adapté à différents types de tâches et de données (Figure I.6). Certains des types de modèles d'apprentissage automatique les plus couramment utilisés incluent :

- **Modèles d'apprentissage supervisé** : ces modèles sont formés sur des données étiquetées, où la sortie correcte est fournie pour chaque exemple dans les données d'entraînement. Le modèle apprend à mapper les données d'entrée à la sortie correcte en fonction des exemples fournis. L'apprentissage supervisé est couramment utilisé pour les tâches de classification et de régression. Sous l'égide de l'apprentissage supervisé tombent : la classification, la régression et la prévision.
    - Classification : dans les tâches de classification, le programme d'apprentissage automatique doit tirer une conclusion à partir des valeurs observées et déterminer à quelle catégorie appartiennent les nouvelles observations. Par exemple, lors du filtrage des e-mails comme "spam" ou "non spam", le programme doit examiner les données d'observation existantes et filtrer les e-mails en conséquence.
    - Régression : dans les tâches de régression, le programme d'apprentissage automatique doit estimer - et comprendre - les relations entre les variables. L'analyse de régression se concentre sur une variable dépendante et une série d'autres variables changeantes, ce qui la rend particulièrement utile pour la prédiction et la prévision.
    - Prévision : La prévision est le processus consistant à faire des prédictions sur l'avenir sur la base des données passées et présentes, et est couramment utilisée pour analyser les tendances.
- **Modèles d'apprentissage non supervisés** : ces modèles sont entraînés sur des données non étiquetées, où la sortie correcte n'est pas fournie pour chaque exemple dans les données d'entraînement. Le modèle doit découvrir par lui-même la structure

sous-jacente des données. L'apprentissage non supervisé est couramment utilisé pour les tâches de regroupement et de réduction de la dimensionnalité.

- **Modèles d'apprentissage par renforcement** : ces modèles apprennent en interagissant avec leur environnement et en recevant des commentaires sous forme de récompenses ou de punitions. Le but de l'apprentissage par renforcement est de maximiser la récompense cumulée reçue par le modèle. Ce type d'apprentissage est couramment utilisé dans des applications telles que les jeux et le contrôle de robots.
- **Apprentissage par transfert**, dans lequel un modèle qui a été entraîné sur une tâche est utilisé comme point de départ pour l'entraînement sur une tâche connexe.
- **Apprentissage profond**, dans lequel un modèle est entraîné à l'aide de réseaux de neurones profonds, ce qui lui permet d'apprendre des relations complexes dans les données.

Ces approches peuvent être utilisées seules ou en combinaison les unes avec les autres, selon le problème à résoudre.

Les modèles d'apprentissage automatique sont entraînés à l'aide de divers algorithmes, qui sont des ensembles d'équations mathématiques qui spécifient comment le modèle doit mettre à jour ses paramètres internes en fonction des données qui lui sont fournies. Certains des algorithmes les plus couramment utilisés pour la formation de modèles d'apprentissage automatique comprennent les arbres de décision, les machines à vecteurs de support et les réseaux de neurones profonds.

Les modèles d'apprentissage automatique ont le potentiel d'être très efficaces pour résoudre des problèmes complexes et faire des prédictions basées sur des données, mais ils ont également certaines limites. L'un des principaux défis de l'apprentissage automatique est que la qualité des prédictions du modèle est aussi bonne que la qualité des données sur lesquelles il est entraîné. De plus, les modèles d'apprentissage automatique peuvent être difficiles à interpréter et à expliquer, ce qui les rend moins transparents que d'autres types d'algorithmes.

I.3.4. Algorithmes d'apprentissage automatique utilisés dans cette thèse

Les algorithmes d'apprentissage automatique sont des algorithmes conçus pour apprendre à partir de données. Ils sont couramment utilisés pour construire des modèles prédictifs capables de faire des prédictions sur des événements futurs sur la base de données précédemment observées.

- **Naïve Bayes Classifier (Apprentissage Supervisé - Classification)**

Le classificateur Naïve Bayes est basé sur le théorème de Bayes et classe chaque valeur comme indépendante de toute autre valeur. Il nous permet de prédire une classe/catégorie, basée sur un ensemble donné de caractéristiques, en utilisant la probabilité. Malgré sa simplicité, le classificateur fonctionne étonnamment bien et est souvent utilisé car il surpasse les méthodes de classification plus sophistiquées.

- **K Means Clustering (apprentissage non supervisé - clustering)**

L'algorithme K Means Clustering est un type d'apprentissage non supervisé, qui est utilisé pour catégoriser des données non étiquetées, c'est-à-dire des données sans catégories ou groupes définis. L'algorithme fonctionne en trouvant des groupes dans les données, avec le

nombre de groupes représentés par la variable K. Il fonctionne ensuite de manière itérative pour attribuer chaque point de données à l'un des K groupes en fonction des caractéristiques fournies.

- **Support Vector Machine (Apprentissage Supervisé - Classification)**

Les algorithmes de support Vector Machine sont des modèles d'apprentissage supervisé qui analysent les données utilisées pour la classification et l'analyse de régression. Ils filtrent essentiellement les données en catégories, ce qui est réalisé en fournissant un ensemble d'exemples de formation, chaque ensemble marqué comme appartenant à l'une ou l'autre des deux catégories. L'algorithme fonctionne ensuite pour construire un modèle qui attribue de nouvelles valeurs à une catégorie ou à l'autre.

- **Logistic Regression (Apprentissage supervisé – Classification)**

La régression logistique se concentre sur l'estimation de la probabilité qu'un événement se produise sur la base des données précédentes fournies. Il est utilisé pour couvrir une variable dépendante binaire, c'est-à-dire où seules deux valeurs, 0 et 1, représentent les résultats.

- **Artificial Neural Networks (apprentissage par renforcement)**

Un réseau de neurones artificiels (ANN) comprend des "unités" disposées en une série de couches, chacune se connectant aux couches de chaque côté. ANNs s'inspirent des systèmes biologiques, tels que le cerveau, et de la façon dont ils traitent l'information. ANNs sont essentiellement un grand nombre d'éléments de traitement interconnectés, travaillant à l'unisson pour résoudre des problèmes spécifiques. ANNs apprennent également par l'exemple et par l'expérience, et ils sont extrêmement utiles pour modéliser des relations non linéaires dans des données de grande dimension ou lorsque la relation entre les variables d'entrée est difficile à comprendre.

- **Decision Tree (apprentissage supervisé – classification/régression)**

Un arbre de décision est une structure arborescente de type organigramme qui utilise une méthode de branchement pour illustrer tous les résultats possibles d'une décision. Chaque nœud de l'arbre représente un test sur une variable spécifique - et chaque branche est le résultat de ce test.

- **Random Forest (Apprentissage Supervisé – Classification/Régression)**

Les forêts aléatoires ou « forêts à décision aléatoire » sont une méthode d'apprentissage d'ensemble, combinant plusieurs algorithmes pour générer de meilleurs résultats pour la classification, la régression et d'autres tâches. Chaque classificateur individuel est faible, mais lorsqu'il est combiné avec d'autres, il peut produire excellents résultats. L'algorithme commence par un « arbre de décision » (un graphique en forme d'arbre ou un modèle de décisions) et une entrée en haut. Il parcourt ensuite l'arborescence, les données étant segmentées en ensembles de plus en plus petits, en fonction de variables spécifiques.

- **K-Nearest-Neighbour (apprentissage supervisé)**

L'algorithme K-Nearest-Neighbour estime la probabilité qu'un point de données soit membre d'un groupe ou d'un autre. Il examine essentiellement les points de données autour d'un seul point de données pour déterminer dans quel groupe il se trouve réellement. Par exemple, si un

point se trouve sur une grille et que l'algorithme essaie de déterminer dans quel groupe se trouve ce point de données (Groupe A ou Groupe B, par exemple), il examinerait les points de données à proximité pour voir dans quel groupe se trouvent la majorité des points.

I.4. Apprentissage profond

L'apprentissage en profondeur est un type d'apprentissage automatique qui consiste à former des réseaux de neurones artificiels sur un grand ensemble de données. C'est ce qu'on appelle l'apprentissage "profond" car les réseaux de neurones comportent de nombreuses couches de neurones artificiels, qui s'inspirent de la structure du cerveau.

Dans l'apprentissage en profondeur, le réseau de neurones apprend à effectuer des tâches en analysant des exemples, plutôt qu'en étant explicitement programmé avec des règles. Cela permet au réseau d'apprendre des modèles complexes et de prendre des décisions basées sur ces modèles.

L'apprentissage en profondeur a été couronné de succès dans un large éventail d'applications, y compris la reconnaissance d'images et de la parole, le traitement du langage naturel et même des jeux comme les échecs. Il a également été utilisé pour des tâches telles que la traduction automatique et les voitures autonomes.

L'un des principaux avantages de l'apprentissage en profondeur est sa capacité à apprendre à partir de données non structurées, telles que des images et du texte. Cela en fait un outil puissant pour les tâches qui nécessitent de comprendre et de traiter de grandes quantités de données non structurées.

Pour former un modèle d'apprentissage en profondeur, un grand ensemble de données est introduit dans le modèle, et le modèle ajuste les poids des neurones artificiels afin de minimiser l'erreur entre la sortie prédite et la sortie correcte. Ce processus est connu sous le nom de l’entraînement du modèle.

Les modèles d'apprentissage en profondeur peuvent être entraînés à l'aide de divers algorithmes, tels que la rétropropagation et la descente de gradient stochastique. Ces algorithmes utilisent un processus appelé optimisation pour trouver les meilleurs poids pour le modèle.

Il existe plusieurs types de réseaux de neurones utilisés dans l'apprentissage en profondeur, notamment les réseaux de neurones convolutifs (CNN) et les réseaux de neurones récurrents (RNN). Les CNN sont particulièrement efficaces pour les tâches de reconnaissance d'images, tandis que les RNN sont utiles pour les tâches impliquant des données séquentielles, telles que la traduction de la langue et la reconnaissance vocale.

L'apprentissage en profondeur est devenu de plus en plus populaire ces dernières années en raison de la disponibilité de grandes quantités de données et du développement de matériel informatique plus puissant. Il a également été aidé par le développement de bibliothèques d'apprentissage en profondeur open source, telles que TensorFlow et PyTorch, qui ont permis aux développeurs de créer et de former plus facilement des modèles d'apprentissage en profondeur.

I.4.1. Approches d'apprentissage en profondeur

Il existe plusieurs approches différentes de l'apprentissage en profondeur, notamment :

- **Réseaux de neurones convolutifs (CNN)** : ils sont utilisés pour les tâches de reconnaissance d'images et de vidéos, ainsi que pour le traitement du langage naturel. Les CNN sont conçus pour traiter des données avec une topologie en forme de grille, telle qu'une image.
- **Réseaux de neurones récurrents (RNN)** : ils sont utilisés pour les tâches qui impliquent des données séquentielles, telles que la traduction de la langue et la reconnaissance vocale. Les RNN sont conçus pour traiter des données à dimension temporelle, comme une série chronologique ou une phrase.
- **Réseaux antagonistes génératifs (GAN)** : ils sont utilisés pour générer de nouvelles données similaires à un ensemble de données donné. Les GAN sont constitués de deux réseaux neuronaux : un réseau générateur et un réseau discriminateur. Le réseau générateur génère de nouvelles données, tandis que le réseau discriminateur tente de distinguer les données générées des données réelles.
- **Auto-encodeurs** : ils sont utilisés pour des tâches telles que la réduction de dimensionnalité et la détection d'anomalies. Les auto-encodeurs sont des réseaux de neurones entraînés à reconstruire leurs données d'entrée en apprenant une représentation compacte des données dans une couche intermédiaire, appelée couche de goulot d'étranglement.

Les approches d'apprentissage en profondeur ont obtenu des résultats de pointe sur un large éventail de tâches, notamment la reconnaissance d'images et de la parole, le traitement du langage naturel et les jeux.

I.4.2. Applications d'apprentissage en profondeur

Reconnaissance d'images et de vidéos : les modèles d'apprentissage en profondeur peuvent être formés pour classer les images et les vidéos dans des catégories prédéfinies, détecter les objets qu'elles contiennent et même générer des descriptions de leur contenu. Ces modèles ont été utilisés dans des applications telles que la reconnaissance faciale, les voitures autonomes et la vidéosurveillance.

- **Traitement du langage naturel** : les modèles d'apprentissage en profondeur peuvent être utilisés pour comprendre et générer le langage humain, permettant des tâches telles que la traduction automatique, la génération de langage et la modélisation du langage.
- **Reconnaissance vocale** : les modèles d'apprentissage en profondeur peuvent être utilisés pour transcrire des mots parlés en texte écrit, permettant des applications telles que la dictée voix-texte et les assistants virtuels.
- **Imagerie médicale** : les modèles d'apprentissage en profondeur peuvent être utilisés pour analyser des images médicales, telles que les rayons X, afin de détecter des anomalies et d'aider au diagnostic.
- **Systèmes de recommandation** : les modèles d'apprentissage en profondeur peuvent être utilisés pour personnaliser les recommandations en fonction de l'historique et des préférences d'un utilisateur. Cela a été appliqué dans des domaines tels que les

recommandations de musique et de films et les suggestions de produits sur les sites Web de commerce électronique.

- **Modélisation financière** : des modèles d'apprentissage en profondeur ont été utilisés pour prédire les cours des actions et analyser les données financières à des fins de gestion des risques et de détection des fraudes.

Ces domaines en dessus ne représente que quelques exemples des nombreuses façons dont l'apprentissage en profondeur peut être appliqué. Les modèles d'apprentissage en profondeur ont la capacité d'apprendre et de prendre des décisions par eux-mêmes, ce qui les rend adaptés à un large éventail de tâches nécessitant un certain degré d'intelligence.

I.5. Web Sémantique

Le Web sémantique est un concept développé par le World Wide Web Consortium (W3C)[1] qui vise à faciliter la compréhension et le traitement par les machines de la signification des informations sur le Web. Il est basé sur l'idée d'utiliser des formats de données normalisés et des métadonnées lisibles par machine pour décrire le contenu et les relations des ressources Web, afin que les ordinateurs puissent plus facilement interpréter et utiliser les informations qu'ils trouvent sur le Web.

Le Web sémantique est construit au-dessus de la structure existante du World Wide Web, en utilisant des technologies telles que RDF (Resource Description Framework) et OWL (Langage d'Ontologie Web) pour ajouter une couche de sens au contenu Web. En utilisant ces normes pour annoter les ressources Web avec des métadonnées lisibles par machine, il devient possible pour les ordinateurs de comprendre les relations et les significations des ressources qu'ils trouvent sur le Web, et d'utiliser ces informations pour effectuer des tâches plus intelligentes telles que faire des inférences et répondre à des questions.

L'objectif du Web sémantique est de permettre aux ordinateurs de comprendre et de traiter la grande quantité d'informations disponibles sur le Web de manière plus significative, et de permettre aux machines de trouver et d'utiliser plus facilement les informations dont elles ont besoin. Il peut nous donner la façon dont nous utilisons le Web et la façon dont nous interagissons avec les ordinateurs, il permet aussi aux machines de nous aider plus facilement dans des tâches telles que la recherche d'informations, la prise de décisions et la résolution de problèmes. Pour atteindre cet objectif, le Web sémantique s'appuie sur un certain nombre de technologies, notamment :

- **RDF**[2] (Resource Description Framework) : Un standard pour décrire la signification des données sur le web, en utilisant un modèle de données simple et flexible.
- **OWL**[3] (Web Ontology Language) : Une norme pour définir la signification des concepts et des relations dans un domaine, en utilisant un langage formel basé sur la logique.
- **SPARQL**[4] : un langage de requête standard pour interroger les données RDF.

[1] https://www.w3.org/
[2] https://www.w3.org/RDF/
[3] https://www.w3.org/OWL/
[4] https://www.w3.org/TR/sparql11-overview/

Ces technologies permettent aux développeurs Web d'annoter leurs données avec une signification supplémentaire, en utilisant un ensemble commun de normes et de langages. Cela permet aux ordinateurs de comprendre la sémantique des données, de les traiter et de les manipuler de manière plus intelligente.

Un aspect clé du Web sémantique est l'idée de données liées, dans laquelle les données sur le Web sont connectées entre elles à l'aide de RDF et d'URL. Cela permet aux données de différentes sources d'être connectées et liées de manière significative, permettant une vue plus interconnectée et intégrée du Web.

Dans l'ensemble, le Web sémantique vise à faire du Web une ressource plus intelligente et utile pour les humains et les ordinateurs, en rendant la signification des informations sur le Web plus explicite et accessible.

I.5.1. Architecture du web sémantique

La pile Web sémantique est une illustration de la hiérarchie des langages, où chaque couche exploite et utilise les capacités des couches inférieures. Il montre comment les technologies standardisées pour le Web sémantique sont organisées pour rendre le Web sémantique possible. Il montre également comment le Web sémantique est une extension (et non un remplacement) du Web hypertexte classique. L'architecture du web sémantique est illustrée dans la Figure I.7.

La première couche, **URI et Unicode**, suit les caractéristiques importantes du WWW existant.

- Unicode est une norme d'encodage des jeux de caractères internationaux et permet que toutes les langues humaines puissent être utilisées (écrites et lues) sur le Web en utilisant une forme normalisée.
- Uniform Resource Identifier (URI) est une chaîne d'un formulaire standardisé qui permet d'identifier de manière unique des ressources (par exemple, des documents). Un sous-ensemble d'URI est :
  - Uniform Resource Locator (URL), qui contient un mécanisme d'accès et un emplacement (réseau) d'un document - tel que http://www.example.org/.

  Un autre sous-ensemble d'URI est :
  - L'URN qui permet d'identifier une ressource sans impliquer son emplacement et les moyens de la déréférencer - un exemple est urn:isbn:0-123-45678-9.

  L'utilisation de l'URI est importante pour un système Internet distribué car elle fournit une identification compréhensible de toutes les ressources. Une variante internationale de l'URI est l'identificateur de ressource internationalisé (IRI) qui permet l'utilisation de caractères Unicode dans l'identificateur et pour lequel un mappage vers l'URI est défini.

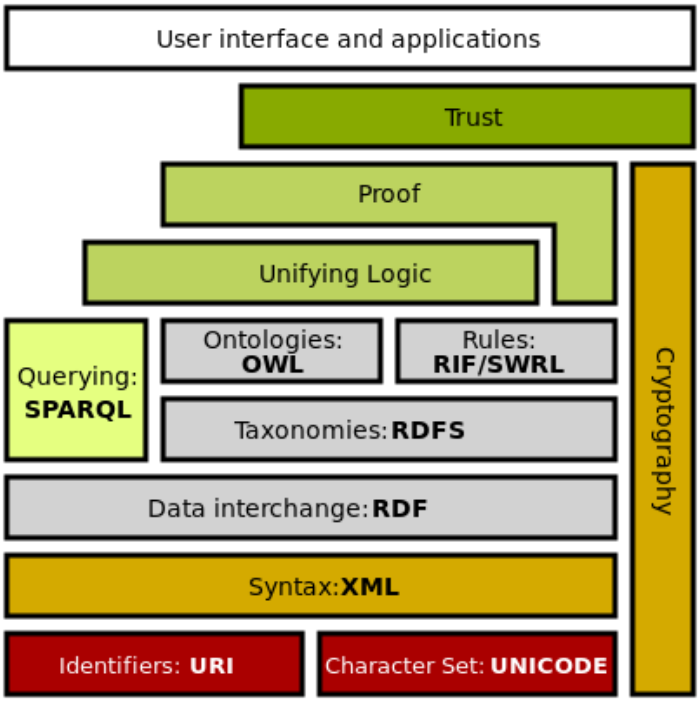

Figure I.7 - Architecture du web sémantique[5].

Dans l'architecture du web sémantique, la couche **XML** (Extensible Markup Language) représente un espace de noms XML et des définitions de schéma XML, elle garantit qu'il existe une syntaxe commune utilisée dans le Web sémantique. XML est un langage de balisage à usage général pour les documents contenant des informations structurées. Un document XML contient des éléments qui peuvent être imbriqués et qui peuvent avoir des attributs et du contenu. Les espaces de noms XML permettent de spécifier différents vocabulaires de balisage dans un document XML. Le schéma XML sert à exprimer le schéma d'un ensemble particulier de documents XML.

Un format de représentation de données de base pour le Web sémantique est le Resource Description Framework (RDF). **RDF** est un Framework pour représenter des informations sur les ressources sous forme de graphe. Il était principalement destiné à représenter des métadonnées sur les ressources WWW, telles que le titre, l'auteur et la date de modification d'une page Web, mais il peut être utilisé pour stocker toute autre donnée. Il est basé sur des triplets sujet-prédicat-objet qui forment un graphe de données. Toutes les données du Web sémantique utilisent RDF comme langage de représentation principal. La syntaxe normative pour la sérialisation de RDF est XML sous la forme RDF/XML. La sémantique formelle de RDF est également définie.

RDF lui-même sert de description d'un graphe formé de triplets. N'importe qui peut définir le vocabulaire des termes utilisés pour une description plus détaillée. Pour permettre une description standardisée des taxonomies et d'autres constructions ontologiques, un schéma RDF (RDFS) a été créé avec sa sémantique formelle au sein de RDF. **RDFS** peut être utilisé pour décrire des taxonomies de classes et de propriétés et les utiliser pour créer des ontologies légères.

[5] https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24)

Des ontologies plus détaillées peuvent être créées avec Web Ontology Language OWL. L'**OWL** est un langage dérivé des logiques de description et offre plus de constructions que RDFS. Il est intégré syntaxiquement dans RDF, donc comme RDFS, il fournit un vocabulaire standardisé supplémentaire. OWL se décline en trois espèces - OWL Lite pour les taxonomies et les contraintes simples, OWL DL pour une prise en charge complète de la logique de description et OWL Full pour une expressivité maximale et une liberté syntaxique de RDF. Puisque OWL est basé sur la logique de description, il n'est pas surprenant qu'une sémantique formelle soit définie pour ce langage.

RDFS et OWL ont une sémantique définie et cette sémantique peut être utilisée pour le raisonnement dans les ontologies et les bases de connaissances décrites à l'aide de ces langages. Pour fournir des règles au-delà des constructions disponibles à partir de ces langages, les langages de règles sont également normalisés pour le Web sémantique.

Pour interroger les données RDF ainsi que les ontologies RDFS et OWL avec des bases de connaissances, un protocole simple et un langage de requête RDF (SPARQL) sont disponibles. **SPARQL** est un langage de type SQL, mais utilise des triplets et des ressources RDF pour faire correspondre une partie de la requête et pour renvoyer les résultats de la requête. Étant donné que RDFS et OWL sont tous deux construits sur RDF, SPARQL peut également être utilisé pour interroger directement des ontologies et des bases de connaissances. Notez que SPARQL n'est pas seulement un langage de requête, c'est aussi un protocole d'accès aux données RDF.

On s'attend à ce que toute la sémantique et les règles soient exécutées au niveau des couches sous « **Preuve »** et le résultat sera utilisé pour prouver les déductions. Une preuve formelle associée à des entrées fiables pour la preuve signifiera que les résultats peuvent être fiables, ce qui est illustré dans la couche supérieure de la Figure I.7. Pour des entrées fiables, des moyens de **cryptographie** sont utilisés, tels que des signatures numériques pour la vérification de l'origine des sources. Au-dessus de ces couches, une **application** avec **interface** utilisateur peut être construite.

I.5.2. Applications du web sémantiques

Le Web sémantique est une vision de l'avenir du World Wide Web, dans laquelle le contenu Web peut être compris et traité par des machines d'une manière plus significative et utile pour les humains. En pratique, cela implique d'utiliser des normes telles que RDF (Resource Description Framework) et OWL (Web Ontology Language) pour annoter le contenu Web avec des métadonnées supplémentaires qui peuvent être utilisées pour décrire la signification et les relations entre différentes informations.

Il existe de nombreuses applications potentielles pour le Web sémantique, notamment :

- **Moteurs de recherche** : en annotant le contenu Web avec des métadonnées supplémentaires, les moteurs de recherche peuvent fournir des résultats de recherche plus pertinents et plus précis.
- **Intégration et interopérabilité des données** : le Web sémantique peut aider à faciliter l'intégration et l'interopérabilité des données provenant de différentes sources, en fournissant un langage commun pour décrire et relier les données.

- **Gestion des connaissances** : Le Web sémantique peut être utilisé pour organiser et gérer les connaissances au sein d'une organisation, en créant une ontologie partagée qui peut être utilisée pour décrire et classer les informations.
- **E-Commerce** : le Web sémantique peut être utilisé pour améliorer l'expérience d'achat en fournissant des informations supplémentaires sur les produits et services, telles que des comparaisons de prix, des avis et des évaluations.
- **Assistants personnels** : le Web sémantique peut être utilisé pour améliorer les capacités des assistants personnels, tels que Siri et Alexa, en leur fournissant une compréhension plus complète de la signification et du contexte des informations qu'ils traitent.
- **Systèmes d'aide à la décision clinique** qui utilisent les technologies du Web sémantique pour analyser et interpréter les données des patients afin de fournir des recommandations de traitement.

I.6. Ontologie

En informatique, l'ontologie est un système d'organisation et de catégorisation de concepts et d'entités dans un domaine de connaissance particulier. Il sert de base pour représenter et raisonner sur les relations entre ces concepts et entités, et peut être utilisé pour faciliter la communication et la compréhension entre les personnes qui utilisent des vocabulaires différents pour décrire les mêmes concepts et entités.

Les ontologies peuvent être utilisées dans diverses applications, telles que le traitement du langage naturel, l'intégration de données et la gestion des connaissances. Ils sont souvent représentés dans un format lisible par machine, tel que le langage d'ontologie Web (OWL), qui permet un traitement et un raisonnement automatisés sur les concepts et les relations définis dans l'ontologie.

Une ontologie consiste généralement en un ensemble de concepts, dont chacun représente une catégorie d'objets ou d'idées, et un ensemble de relations entre ces concepts. Les concepts et les relations d'une ontologie sont souvent organisés en hiérarchie, avec des concepts plus généraux en haut et des concepts plus spécifiques en bas.

Par exemple, une ontologie pour un domaine tel que la biologie peut inclure des concepts tels que « animal », « plante » et « organisme », ainsi que des relations telles que « est un sous-type de » et « fait partie de ». Cette ontologie pourrait ensuite être utilisée pour représenter et raisonner sur les relations entre différents types d'animaux, de plantes et d'organismes.

Les ontologies sont un outil important dans le domaine de l'intelligence artificielle et sont souvent utilisées pour prendre en charge des tâches telles que la recherche d'informations, l'apprentissage automatique et le traitement du langage naturel. Ils peuvent également être utilisés pour faciliter l'interopérabilité entre différents systèmes et bases de données en fournissant un vocabulaire partagé pour décrire et organiser les données.

I.6.1. Composants de l’ontologie

Plusieurs composants sont généralement inclus dans une ontologie :

- **Classes** : ce sont les principaux concepts ou catégories du domaine. Chaque classe représente un ensemble d'objets qui partagent des caractéristiques et des propriétés communes.
- **Propriétés** : elles décrivent les caractéristiques et les attributs des objets du domaine. Les propriétés peuvent être simples (par exemple, " name", " age") ou complexes (par exemple, " has child").
- **Relations** : elles décrivent les connexions et les associations entre différentes classes et propriétés dans l'ontologie. Les relations peuvent être hiérarchiques (par exemple, " subclass of"), associatives (par exemple, " part of") ou fonctionnelles (par exemple, " has value").
- **Instances** : ce sont les objets ou entités spécifiques qui appartiennent à une classe particulière dans l'ontologie.
- **Axiomes** : ce sont des déclarations qui définissent les relations et les contraintes au sein de l'ontologie. Les axiomes peuvent être utilisés pour spécifier des informations supplémentaires sur les classes, les propriétés et les relations dans l'ontologie.
- **Annotations** : Ce sont des informations supplémentaires qui peuvent être associées aux concepts et aux relations dans l'ontologie. Les annotations peuvent inclure des définitions, des exemples et d'autres métadonnées descriptives.

Les ontologies sont utilisées pour fournir un langage et une compréhension communs d'un domaine, et peuvent être utilisées pour faciliter la communication et l'échange de données entre différents systèmes et applications.

I.6.2. Classifications d'ontologies

Les ontologies sont des représentations formelles et explicites des concepts et des relations au sein d'un domaine de connaissances particulier. Il existe plusieurs façons de classer les ontologies, notamment les suivantes :

Objectif : les ontologies peuvent être classées en fonction de leur objectif ou de l'utilisation prévue de l'ontologie. Par exemple, certaines ontologies sont conçues pour représenter des connaissances à utiliser dans des tâches de traitement du langage naturel, tandis que d'autres sont conçues pour prendre en charge l'intégration de données ou l'interopérabilité entre différents systèmes.

Structure : les ontologies peuvent également être classées en fonction de leur structure ou de la manière dont elles représentent les connaissances. Une façon courante de classer les ontologies en fonction de la structure consiste à faire la distinction entre les ontologies taxonomiques, qui représentent une hiérarchie de concepts et de relations, et les ontologies assertionnelles, qui représentent un ensemble d'énoncés sur les relations entre les concepts.

Formalité : les ontologies peuvent également être classées en fonction du niveau de formalité ou de rigueur avec lequel elles sont développées. Certaines ontologies sont très formelles et suivent des directives strictes pour représenter les connaissances, tandis que d'autres sont plus flexibles et peuvent être développées de manière informelle.

Domaine : les ontologies peuvent également être classées en fonction du domaine de connaissances qu'elles représentent. Par exemple, il peut y avoir des ontologies pour le domaine de la biologie, le domaine de la finance ou le domaine des médias sociaux.

Langage : les ontologies peuvent également être classées en fonction du langage ou du formalisme de représentation utilisé pour représenter les concepts et les relations au sein de l'ontologie. Par exemple, certaines ontologies utilisent le Web Ontology Language (OWL), tandis que d'autres utilisent le Resource Description Framework (RDF) ou d'autres formalismes de représentation.

I.6.3. Structure de l'ontologie

Les ontologies peuvent être catégorisées selon leur formalité, qui détermine le degré d'axiomatisation des instructions logiques. De nombreuses approches du Web sémantique ont été utilisées pour la modélisation de domaine, telles que le thésaurus, la taxonomie, les modèles conceptuels, etc.

- **Thésaurus** : ce modèle est utilisé pour organiser les termes de la connaissance d'un domaine spécifique avec des restrictions aux relations lexicales telles que l'homonyme et le synonyme. WordNet est un exemple bien connu de modèle de thésaurus.
- **Taxonomie** : ce modèle représente la structure formelle des classes ou types d'objets au sein d'un connaissance du domaine. La taxonomie est une méthode de catégorisation des termes de vocabulaire dans une structure hiérarchique. La racine du modèle hiérarchique est le concept général de l'arbre. Les nœuds de l'arbre représentent les termes avec une connexion à d'autres nœuds via des relations parent/enfant. Par conséquent, les machines apprennent efficacement à l'aide de taxonomies et peuvent faire des inférences statistiques et des associations statistiques basées sur la proximité.
- **Modèles conceptuels** : ces modèles sont utilisés pour exprimer la structure de données de la connaissance du domaine au moyen de classes, d'attributs et de relations tels que le langage de modélisation unifié (UML) et le diagramme de relation d'entité (ERD).

I.6.4. Principaux rôles de l'ontologie

En général, l'ontologie fait référence à l'étude de la nature de l'être, du devenir, de l'existence ou de la réalité, ainsi qu'aux catégories et principes de base qui sous-tendent l'organisation et la construction des connaissances. Dans le domaine de l'informatique et des technologies de l'information, les ontologies sont utilisées pour représenter et organiser les connaissances de manière structurée et formalisée, permettant aux machines de comprendre et d'interpréter le sens et les relations entre différents concepts et entités.

Les ontologies peuvent jouer plusieurs rôles principaux dans divers contextes :

1. **Représentation des connaissances** : les ontologies fournissent un moyen de représenter les connaissances de manière structurée, formalisée et lisible par machine. Ils peuvent être utilisés pour décrire les concepts, les relations et les hiérarchies qui existent dans un domaine de connaissances particulier, permettant aux machines de comprendre et de raisonner sur ces connaissances.
2. **Intégration et interopérabilité des données** : les ontologies peuvent être utilisées pour permettre l'intégration et l'utilisation conjointe de données provenant de différentes sources, en fournissant un vocabulaire commun et un ensemble de concepts pouvant être utilisés pour décrire les données. Cela permet une

communication et une collaboration plus efficaces entre les différents systèmes et organisations.

3. **Extraction et recherche d'informations** : les ontologies peuvent être utilisées pour améliorer la précision et l'efficacité des systèmes d'extraction et de recherche d'informations en fournissant une représentation plus précise et structurée des connaissances recherchées.
4. **Traitement du langage naturel** : les ontologies peuvent être utilisées pour améliorer les performances des systèmes de traitement du langage naturel en fournissant une représentation structurée des concepts et des relations mentionnés dans le texte. Cela peut aider les systèmes NLP à comprendre le sens des mots et des phrases dans leur contexte, et à générer des réponses plus précises et pertinentes.
5. **Intelligence artificielle et apprentissage automatique** : les ontologies peuvent être utilisées pour soutenir le développement et l'application de systèmes d'intelligence artificielle (IA) et d'apprentissage automatique (ML), en fournissant une représentation structurée des concepts et des relations qui sont pertinents pour un domaine ou une tâche particulière. Cela peut aider les systèmes d'IA et de ML à apprendre plus efficacement et à prendre des décisions plus éclairées.

I.6.5. Apprentissage automatique basé sur des ontologies

L'apprentissage automatique basé sur des ontologies est un sous-domaine de l'intelligence artificielle qui se concentre sur l'utilisation d'ontologies (représentations formelles de concepts et de leurs relations au sein d'un domaine) pour améliorer les performances des algorithmes d'apprentissage automatique.

En général, les approches basées sur l'ontologie de l'apprentissage automatique impliquent l'intégration de connaissances spécifiques à un domaine dans le processus d'apprentissage afin d'améliorer la précision et l'interprétabilité des modèles résultants. Cela peut se faire de plusieurs manières, notamment en incorporant des contraintes spécifiques au domaine dans l'algorithme d'apprentissage, en prétraitant les données pour les rendre plus propices à l'apprentissage, ou en utilisant l'ontologie pour guider le processus d'apprentissage lui-même.

Une application courante de l'apprentissage automatique basé sur l'ontologie est la modélisation prédictive, où l'objectif est d'utiliser des données historiques pour faire des prédictions sur des événements futurs. En incorporant des connaissances spécifiques au domaine dans l'algorithme d'apprentissage, les approches basées sur l'ontologie peuvent aider à améliorer la précision des modèles résultants, leur permettant de faire des prédictions plus précises.

Par exemple, une approche basée sur l'ontologie peut être utilisée pour prédire la probabilité qu'un client effectue un achat en fonction de son comportement passé. En incorporant des connaissances sur les préférences et l'historique d'achat du client, ainsi que des connaissances sur les produits proposés, un algorithme d'apprentissage automatique basé sur une ontologie pourrait être en mesure de faire des prédictions plus précises qu'un algorithme d'apprentissage automatique traditionnel qui ne prend pas ces connaissances en compte.

Dans l'ensemble, la recherche dans le domaine de l'apprentissage automatique basé sur l'ontologie se concentre sur le développement d'algorithmes et de techniques qui peuvent efficacement intégrer des connaissances spécifiques à un domaine pour améliorer les

performances des modèles d'apprentissage automatique, avec un accent particulier sur les applications en modélisation prédictive.

I.6.6. Raisonnement et inférence ontologique

L'un des principaux avantages des ontologies est qu'elles permettent l'intégration du raisonnement et de l'inférence dans le modèle. Cela signifie qu'une fois qu'une ontologie est créée, elle peut être utilisée pour déduire automatiquement de nouvelles informations basées sur les relations et les concepts définis dans l'ontologie.

Les ontologies sont souvent construites à l'aide du langage OWL, qui est un langage Web sémantique conçu spécifiquement pour la création d'ontologies. OWL a une sémantique formelle qui définit la signification des concepts et des relations définis dans une ontologie, et cette sémantique est utilisée par les moteurs de raisonnement pour déduire automatiquement de nouvelles informations basées sur l'ontologie.

Des mécanismes de raisonnement sont implémentés dans des moteurs d'inférence, qui sont des logiciels conçus pour déduire automatiquement de nouvelles informations à partir de règles et de concepts définis dans une ontologie. Ces moteurs permettent de déterminer certaines actions liées à l'ontologie et à sa hiérarchie, et d'en déduire des conséquences logiques en fonction des classes définies dans l'ontologie. Un exemple de langage de règles pouvant être utilisé dans les moteurs d'inférence est SWRL (plus de détail sur ce langage dans la section suivante), qui permet la création de règles pouvant être exécutées automatiquement par les moteurs d’inférence.

I.7. Langage de règles du web sémantique

Le langage de règles du web sémantique (SWRL) est un langage de règles pour le Web sémantique[6]. Il est conçu pour faciliter l'intégration de données provenant de différentes sources sur le Web sémantique en permettant de spécifier des règles qui relient différentes sources de données les unes aux autres. SWRL est basé sur une combinaison du langage d'ontologie Web (OWL-DL) et du langage de balisage de règles (RuleML « Rule Markup Language »).

Les règles SWRL ont la forme « si *condition* alors *conclusion* », où la condition est un ensemble d'atomes et la conclusion est un ensemble d'atomes. Les atomes sont constitués d'un prédicat (une propriété ou une relation) et d'un ensemble d'arguments (des individus ou des ressources). Les atomes de la condition sont appelés l'**antécédent** de la règle et les atomes de la conclusion sont appelés le **conséquent**. La première partie (Antécédent) spécifie les conditions qui doivent être vérifiées et la deuxième partie (Conséquent) spécifie les actions à faire.

Les règles SWRL peuvent être utilisées pour spécifier des relations entre différentes sources de données, ainsi que pour effectuer des inférences et des dérivations basées sur ces relations. Par exemple, une règle SWRL peut spécifier que si une personne est membre d'une certaine organisation, elle est également membre d'un groupe particulier au sein de cette organisation. Cette règle pourrait être utilisée pour classer automatiquement les individus en fonction de leur appartenance à différentes organisations.

[6] https://www.w3.org/Submission/SWRL/

Les moteurs d'inférence sont des outils logiciels qui peuvent raisonner sur un ensemble d'énoncés ou de règles et tirer de nouvelles conclusions sur la base de ces informations. Il existe plusieurs moteurs d'inférence qui sont compatibles avec SWRL et peuvent être utilisés pour raisonner sur les règles SWRL. Voici quelques exemples de moteurs d'inférence prenant en charge SWRL :

- **Pellet** : est un moteur de raisonnement open source pour les ontologies OWL. Il prend en charge un large éventail de fonctionnalités OWL et inclut la prise en charge des règles SWRL.
- **HermiT** : est un moteur de raisonnement open source pour les ontologies OWL. Il est conçu pour être très efficace et évolutif, et inclut la prise en charge des règles SWRL.
- **RacerPro** : est un moteur de raisonnement commercial pour les ontologies OWL. Il inclut la prise en charge des règles SWRL et est conçu pour être rapide et évolutif.
- **Fact++** : est un moteur de raisonnement open-source pour les ontologies OWL. Il inclut la prise en charge des règles SWRL et est conçu pour être efficace et évolutif.

Il existe de nombreux autres moteurs d'inférence qui prennent en charge SWRL et peuvent être utilisés pour raisonner sur les règles SWRL.

SWRL fait partie de la pile Web sémantique plus large, qui comprend des technologies telles que RDF, OWL et SPARQL. Ces technologies sont conçues pour permettre de combiner et d'interroger facilement des données provenant de différentes sources, afin de faciliter la création de systèmes plus intelligents et plus efficaces.

I.8. Conclusion

Dans ce chapitre, nous avons présenté les piliers techniques de cette thèse, qui sont issus de deux domaines le web sémantique et l'intelligence artificielle, plus spécifiquement les ontologies et l'apprentissage automatique. Ils caractérisent des outils fréquemment appliqués à la prédiction dans le domaine de la santé en particulier, et dans d'autres secteurs en générale.

En premier lieu, nous avons exploré les notions de base liées aux techniques de l'intelligence artificielle à savoir l'apprentissage automatique et l'apprentissage profond en donnant leurs architectures et leurs modes de fonctionnement et d'application.

En deuxième lieu, nous avons donné un aperçu du web sémantique de son architecture et de ses applications, et nous avons exposé aussi une vue générale sur les ontologies en détaillant ses caractéristiques fondamentales, les composants de l'ontologie, la structure de l'ontologie, l'apprentissage automatique basé sur des ontologies et nous avons clôturé cette partie par le raisonnement et inférence ontologique.

Dans le chapitre suivant, nous exposons un état de l'art exhaustif sur l'emploi des ontologies et des techniques d'apprentissage automatique pour la prédiction des maladies cardiovasculaires, le cancer du sein, et la détection des cas du COVID-19. Nous allons voir également l'impact d'appliquer l'apprentissage automatique basé sur des ontologies pour avoir de meilleurs résultats.

CHAPITRE II - REVUE DE L'ETAT DE L'ART

II.1. Introduction

La science des données est un domaine multidisciplinaire qui utilise des méthodes, des processus, des algorithmes et des systèmes scientifiques pour extraire des connaissances et des idées à partir de données structurées et non structurées [4]. Les statistiques, l'exploration de données, la visualisation de données, l'apprentissage automatique, l'apprentissage en profondeur et l'intelligence artificielle sont les principaux sous-thèmes de la science des données. Même si la science des données est née dans les années 1990, l'importance de ce domaine se réalise de nos jours. Il est mentionné dans différentes études que la quantité de données dans le monde augmente rapidement et que le type de données non structurées représente toujours plus de la moitié de la quantité totale de données. Par conséquent, la science des données est devenue un enjeu essentiel dans tous les domaines pour rendre les données compréhensibles. La santé est l'un des environnements nécessaires aux applications de la science des données puisque le big data en fait partie. Le volume de données collectées dans le domaine de la santé est énorme, pourtant il est prouvé que 80% des données collectées ne sont pas organisées. Le nombre total d'études parmi les applications de la science des données dans le domaine de la santé a considérablement augmenté.

L'environnement de la santé est l'un des domaines les plus précis pour les applications de science des données en raison de la quantité de données qu'il contient et de la pertinence du type de données. Le flux de données dans les hôpitaux est un processus continu et comprend des valeurs numériques en général. Healthcare est un système ouvert d'amélioration avec des études sur l'exploration de données et les techniques d'apprentissage automatique. Les chercheurs affirment que l'expertise sur un ordinateur vous donnerait des résultats significatifs et la possibilité de prédire l'avenir avec l'historique des données passées [5]. De nombreuses études ont été réalisées sur des ensembles de données sur des maladies différentes, et la plupart d'entre elles ont une précision de classification suffisante.

Dans cette revue de la littérature, un aperçu général de la littérature existante sur la prédiction des maladies dans le domaine de la santé à l'aide de l'apprentissage automatique et de l'ontologie est donné. Cet aperçu est volontairement très général. Dans les chapitres suivants de cette thèse, une revue de littérature plus spécifique est donnée dans chaque chapitre technique.

II.2. Méthodes de prédiction basées sur l'apprentissage automatique

Avec les progrès de la technologie, l'apprentissage automatique devient une technologie de plus en plus populaire et couramment utilisée par les experts de l'industrie pour résoudre les problèmes rencontrés dans la vie réelle. L'apprentissage automatique est l'étude scientifique des algorithmes et des modèles statistiques que l'ordinateur utilise pour effectuer une tâche spécifique sans utiliser d'instructions explicites, en s'appuyant plutôt sur des modèles et des inférences. L'apprentissage automatique est également utilisé par l'industrie de la santé pour faire progresser leurs techniques afin qu'ils puissent fournir de meilleurs services à leurs patients. Le système de prédiction des maladies prédit les maladies en fonction des

symptômes du patient et également de certains médicaments couramment prescrits pour une maladie particulière.

De nombreux travaux de recherche ont été menés pour prédire les maladies en fonction des symptômes présentés par un individu à l'aide des algorithmes d'apprentissage automatique. Nous en citons des travaux liés au cancer du sein, cardiovasculaire et COVID-19, dans les sous-sections suivantes.

II.2.1. Détection du cancer du sein à l'aide de l'apprentissage automatique

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle (IA) qui utilise le codage logiciel au lieu de l'approche conventionnelle codée. L'apprentissage automatique fait référence au mécanisme par lequel, malgré l'absence d'instructions explicitement programmées, la machine peut continuer à apprendre de l'expérience [6], [7]. Dans la recherche sur le cancer, les modèles d'apprentissage automatique ont une longue histoire d'être utiles non seulement pour la recherche, mais aussi pour la mise en œuvre pratique dans la détection du cancer [8]. Depuis près de 30 ans ont passé, les arbres décisionnels et les réseaux de neurones artificiels contribuent à la détection et au diagnostic du cancer. Près de 20 ans se sont écoulés depuis l'introduction des modèles basés sur la SVM en tant que modèle pronostique du cancer. Plusieurs systèmes optimisés ont été introduits pour le traitement d'images dans le secteur médical tels que les systèmes CADe et CADx à l'aide de plusieurs algorithmes basés sur l'apprentissage automatique. Les systèmes CADe aident à détecter les objets qui ont une grande importance en termes cliniques, tandis que les systèmes CADx aident à quantifier la malignité des objets cliniques qui sont détectés manuellement ou automatiquement [9], [10]. Divers algorithmes et techniques d'exploration de données ont été appliqués dans plusieurs études menées sur une multitude d'ensembles de données pour classer le cancer du sein. Le principal avantage de ces techniques et algorithmes est leurs résultats de classification supérieurs. Cela a conduit de nombreux chercheurs à intégrer des techniques telles que l'exploration de données et l'utilisation de modèles d'apprentissage automatique optimisés dans leurs études qui tournent autour de la résolution de tâches difficiles et alambiquées [11]. La Table II.1 donne un aperçu des articles qui utilisent des approches d'apprentissage automatique appliquées au cancer du sein.

Les chercheurs ont également travaillé sur l'augmentation de la précision de la prédiction du temps de survie pour une personne diagnostiquée avec un cancer du sein, en utilisant l'apprentissage automatique [12]. Essentiellement, cette étude tente d'évaluer l'efficacité et la précision des algorithmes d'apprentissage automatique déjà en place pour prédire le temps de survie. Les auteurs ont proposé une approche utilisant l'apprentissage automatique et ajouté une nouvelle fonctionnalité basée sur la concaténation de trois caractéristiques, à savoir le stade de la tumeur, la taille de la tumeur et l'âge au moment du diagnostic, en une seule nouvelle fonctionnalité. Ensuite, ils ont appliqué des modèles d'apprentissage automatique à l'aide de Support Vector Machine (SVM)- Regression (SVR). Les méthodes utilisées par eux ont fourni des résultats optimistes. Les auteurs ont également montré que des prédictions plus précises pouvaient être faites en utilisant à la fois des modèles de régression linéaires et basés sur des arbres de décision de SVR et ont confirmé la même chose en utilisant la validation croisée. Enfin, ils ont conclu que la nouvelle fonctionnalité clinique intégrée de la tumeur (TCIF) surpassait la fonctionnalité existante de l'indice pronostique de Nottingham (NPI).

Les auteurs [13], ont proposé un classificateur croisé basé sur l'approche de découverte de connaissances pour détecter la présence d'un cancer du sein. L'objectif de l'article était d'effectuer une analyse approfondie et de fournir une comparaison des différents programmes algorithmiques d'apprentissage automatique tels que SVM, Naive Bayes, K-Nearest Neighbours et DT. Les WBCD ont été utilisés pour implémenter les algorithmes. Leur objectif principal était d'évaluer les performances des algorithmes sur plusieurs paramètres afin de développer un nouvel algorithme de fusion qui afficherait une exécution optimale. Sur la base de leurs expériences, les auteurs ont découvert qu'un classificateur qui fusionnait trois types de modèles de SVM, NB et C4.5 pouvait atteindre une exactitude de 97,31 %, ce qui était le plus performant. Ils ont montré que cette nouvelle approche consistant à utiliser des multi-classificateurs était assez efficace et fiable dans la prédiction du cancer du sein et son diagnostic. Ils ont également conclu que la construction d'un classificateur précis et efficace sur le plan informatique est un véritable défi étant donné que la vie des patients en dépend et doit donc être surveillée attentivement.

Les auteurs de cette étude [14], ont utilisé un modèle, basé sur les variables du noyau cellulaire, qui avait été formé à l'aide de l'apprentissage automatique. Les algorithmes à l'œuvre dans cette étude étaient K-NN et SVM. Les performances de leurs classificateurs ont été déterminées et analysées. À l'aide d'un réseau bayésien, ils visaient à comparer et à contraster la tâche à l'aide d'un ensemble de données contenant les valeurs de caractéristiques recueillies à partir des images de lames de cellules à l'aide de la FNAC. Leurs recherches ont montré que toutes les tentatives visaient à développer un algorithme permettant de prédire si la tumeur est maligne ou bénigne. L'image fournie avait un éclairage variable à divers endroits qui différait en fonction de l'intensité de l'éclairage. Trois niveaux de classification ont été appliqués à l'image : noir, blanc et gris. Le modèle a montré un niveau d'efficacité significativement élevé (97,49%). Par conséquent, les auteurs ont conclu qu'un diagnostic plus rapide du cancer du sein pourrait être réalisé à l'avenir en utilisant un processus par lequel les images de diapositives des cellules obtenues par FNAC pourraient être prétraitées en utilisant l'automatisation pour extraire les caractéristiques pertinentes et être immédiatement introduites directement dans le modèle basé sur l'apprentissage automatique.

Sur la base de l'ensemble de données « Wisconsin Breast Cancer » et en mesurant les performances des onze algorithmes d'apprentissage automatique utilisés pour la tâche de classification, une analyse comparative entre eux a été présentée par [15]. Pour différencier les masses mammaires bénignes et malignes, les auteurs ont proposé une méthode pour développer deux classificateurs en utilisant les caractéristiques dérivées des images post-diagnostic de la FNAC. Ils visaient à examiner et à analyser la précision des 11 algorithmes d'apprentissage automatique différents puis précise l'algorithme qui donne le meilleur résultat. Les résultats des expériences ont montré que le réseau de neurones s'est avéré être le plus précis parmi tous les autres avec un indice d'exactitude de 96,49 %, suivi de l'analyse discriminante linéaire (LDA) et de la régression logistique (LR), puis de la forêt aléatoire (RF), arbre de décision (DT) et classificateur de vecteurs de support (SVC) (linéaire). Ils ont conclu qu'à l'avenir, l'algorithme de réseau neuronal pourrait être déployé et implémenté dans un test Big Data qui pourrait mieux fonctionner que les modules actuels comme Hadoop et Apache Spark.

Dans l'article [16] Les auteurs ont collecté des données à partir de 699 échantillons dans le référentiel d'apprentissage automatique de l'UCI. Le réseau de neurones artificiels (ANN)

utilisé pour cette étude a été configuré de manière à pouvoir utiliser neuf neurones (nombre d'attributs) et un seul neurone de sortie (la nature de la masse, c'est-à-dire bénigne ou maligne). Une matrice de confusion spécifique au réseau de rétropropagation a également été présentée. Les résultats ont indiqué qu'il y avait 99 % de chances d'obtenir un diagnostic correct si l'ANN était fonctionnellement déployé. Cela indiquait également qu'il y avait 97,6% de chances qu'ils soient classés négativement. L'ANN a encore de la place pour d'autres améliorations à l'avenir. Cela pourrait aider les médecins avec une méthode de diagnostic plus rapide et également surveiller l'état du patient. Les auteurs ont proposé de développer un outil basé sur une interface utilisateur graphique (GUI) pour faciliter l'utilisation par les médecins qui ne sont pas particulièrement aptes à utiliser de tels outils sans une interface appropriée.

Dans cette étude [17], les auteurs ont présenté quatre algorithmes basés sur des modèles d'apprentissage automatique pour la détection du cancer du sein : réseaux bayésiens, kNN, SVM et forêt aléatoire. Les auteurs visaient à développer une approche méticuleuse pour diagnostiquer et classer le cancer du sein. Contrairement à la méthode de la forêt aléatoire, la méthode Support Vector Machine (SVM) a atteint une précision et un caractère distinctif optimaux, la première ayant la plus grande probabilité de classer les tumeurs de manière appropriée. À l'aide de propriétés telles que le temps de formation pour un petit et un grand ensemble de données, le rappel, l'aire sous la courbe caractéristique de fonctionnement du récepteur, la précision des prédictions faites et le cumul du nombre de caractéristiques, une comparaison précise peut être établie. Leur étude a montré qu'il y a encore de la place pour des futures investigations dans ce domaine.

Les auteurs de cet article [18] ont utilisé l'ensemble de données WBCD et ont présenté une comparaison directe entre cinq algorithmes de diagnostic du cancer du sein : Perceptron multicouche, kNN, Arbres de classification et de régression, NB et SVM. Tout au long de l'étude, l'objectif principal était d'évaluer les performances des différentes méthodes d'apprentissage automatique en évaluant la fiabilité, l'exactitude et la précision de chaque algorithme individuellement pour la classification des données. L'objectif était de découvrir quelle méthode d'apprentissage automatique pourrait prédire la nature de la tumeur. Leurs résultats ont montré que le modèle Multilayer Perceptron offre des performances optimales en termes de précision, de rappel et d'exactitude. L'exactitude montrée par le MLP sur les données d'entraînement était de 96,70 %, ce qui surpassait les autres algorithmes. Ces modèles ont ensuite été testés sur des données récentes pour analyser leurs performances dans le monde réel.

Dans le papier [19], les auteurs ont présenté quatre algorithmes de classification différents avec des comparaisons et des contrastes entre eux : Support Vector Machine, Naive Bayes, kNN et arbres de décision C4.5. Sur la base des mesures de performance de la sensibilité, de la précision, de la spécificité et de l'exactitude, ils ont cherché à évaluer l'efficience et l'efficacité des algorithmes. Le logiciel Weka a été utilisé pour dériver tous les classificateurs employés. Le WBCD a été utilisé dans leur étude qui comprenait 699 cas, dont 458 bénins et 241 malins, répartis en deux classes et 11 caractéristiques, avec 65,5% de tumeurs malignes et 34,5% de tumeurs bénignes. Ils ont appliqué une méthode connue sous le nom de validation croisée K-Folds. La valeur de k a été sélectionnée à 10 pour obtenir le moins de biais. Les résultats du modèle ont ensuite été évalués à l'aide de la validation croisée K-Folds. Ces chercheurs ont évalué les résultats de performance en termes de cinq mesures : les

statistiques Kappa, l'erreur absolue moyenne, l'erreur quadratique moyenne, l'erreur absolue relative et l'erreur quadratique relative racine. Sur la base des résultats de leur expérience, ils ont constaté que le modèle basé sur SVM pouvait atteindre les meilleures performances avec une précision de 97,13 % et que le taux d'erreur le plus faible était de 0,02 %. Les performances du modèle d'apprentissage automatique basé sur d'autres algorithmes variaient entre 95,12 % et 95,28 %, et le taux d'erreur variait entre 0,03 et 0,06. Ils ont observé un grand nombre de classifications incorrectes avec l'algorithme C4.5 et pour l'algorithme k-NN (respectivement 34 et 33 instances incorrectes).Les auteurs de cet article [20], ont rapporté une comparaison de deux techniques d'apprentissage automatique, l'algorithme de réseau bayésien et J48. Ils ont principalement travaillé avec le jeu de données WBCD et ont développé deux classificateurs pour distinguer les lésions bénignes des lésions malignes. Le prétraitement des données est une étape importante avant l'exécution de l'algorithme, car il ne serait pas en mesure de traiter les valeurs manquantes autrement. De plus, les performances du modèle seraient améliorées par un apprentissage supervisé. De plus, en diminuant le nombre de valeurs à l'intérieur des données continues à l'aide du codage d'étiquette, l'efficacité du modèle serait également augmentée. Le logiciel Weka 3.6 a été utilisé pour tous les tests qui ont été effectués. Les performances optimales ont été atteintes lorsque huit attributs ont été utilisés en plus de la classe contenant les valeurs manquantes supprimées. L'algorithme des réseaux bayésiens, qui a atteint des performances optimales parmi tous les algorithmes testés dans cet article, a pu atteindre une précision de 97,80 % dans sa meilleure configuration.

Les auteurs [21] ont proposé une nouvelle approche d'apprentissage automatique kNN pour détecter plus précisément le cancer du sein. Ils ont conçu une méthode composée principalement de deux parties : la première partie de l'approche traitait les images reçues en entrée pour l'extraction de caractéristiques, tandis que la deuxième partie consistait à extraire les caractéristiques qui ont été traitées à l'aide des deux modèles, dont l'un était un réseau de neurones et l'autre modèle employait la régression logistique. De plus, ils ont comparé les deux modèles et les ont analysés à l'aide du logiciel Matlab. Au cours de l'étude, la valeur d'erreur signalée était <0,07 et le modèle basé sur le réseau neuronal utilisait un nombre inférieur de fonctionnalités par rapport à celui qui utilisait LR. Bien que le nombre de fonctionnalités consommées par le modèle BPNN soit inférieur de 24 à celui du modèle basé sur LR, ils ont pu obtenir un taux de réussite > 93 % en utilisant ce modèle.

Les auteurs de l’étude [22]ont présenté leurs travaux sur l'utilisation d'une approche basée sur l'apprentissage automatique pour le diagnostic précoce du cancer du sein, dans laquelle ils ont utilisé des méthodes d'apprentissage automatique pour analyser le nombre de patients cancéreux atteints de tumeurs et présenter un rapport à ce sujet. Il y avait 567 lignes de données représentant 30 attributs distincts des traits du cancer du sein dans l'ensemble de données sur le cancer du sein utilisé par les auteurs. Les lignes contenant des informations sur la nature de la tumeur (c'est-à-dire bénigne ou maligne) ont été retirées et définies comme attributs cibles. Sur la base des données recueillies, un pourcentage a été calculé pour représenter les patients atteints de tumeurs, et les prédictions qui ont été présentées à l'aide de visualisations illustratives. Pour obtenir des estimations impartiales, les chercheurs ont utilisé des algorithmes d'apprentissage automatique supervisé. Sur la base des résultats des tests, ils ont trouvé que les K-Nearest Neighbors étaient le prédicteur le plus précis avec une précision de 91,6 %. La moindre précision a été affichée par l'approche NB, qui a montré une précision

de 75,6 %, tandis que l'approche basée sur kNN a montré une précision de 90,9 %, qui était la plus élevée.

Dans le papier [23], les auteurs ont présenté une analyse comparative de l'algorithme RVM, qui a entraîné des coûts de calcul considérablement inférieurs par rapport à d'autres algorithmes également utilisés pour le diagnostic du cancer du sein. L'étude a évalué comment l'approche RVM offrait un avantage pour diagnostiquer correctement le cancer du sein même lorsque les fonctionnalités étaient réduites par rapport à d'autres modèles d'apprentissage automatique. Le RVM a été comparé à des algorithmes d'apprentissage automatique tels que Naive Bayes, les réseaux de neurones, DT, SVM et les systèmes d'inférence floue pour l'analyse des performances. L'étude a montré que RVM fonctionnait nettement mieux que les autres modèles. Des études antérieures ont montré qu'une approche basée sur la RVM avait rarement été utilisée pour l'ensemble de données WBCD pour le diagnostic du cancer du sein, mais était plus largement acceptée pour d'autres types de cancers tels que le cancer du sang et le cancer lymphatique. Par conséquent, les auteurs ont utilisé le WBCD original dans son étude pour détecter le cancer du sein, qui surpassait toutes les autres approches de l'époque et affichait une précision de 97 %. Même lorsque les fonctionnalités (variables) ont été réduites, RVM a toujours montré de meilleures performances que les autres. Pourtant, l'étude a déclaré qu'il y avait des possibilités d'amélioration à l'avenir et a suggéré qu'elle pourrait également être fusionnée avec d'autres algorithmes d'apprentissage automatique pour augmenter encore la précision par un réglage fin.

Les expériences présentées dans l'étude [24], ont été discernées en utilisant un test de validation croisée k-fold. Dans la première étape, les données d'entrée ont été divisées au hasard en cinq parties et pour chaque calcul, une partie a été utilisée comme ensemble de données de test et les autres quatre parties ont été marquées comme ensemble de données d'apprentissage (répartition 80:20). L'avantage le plus important de la méthode était qu'elle n'était pas pertinente pour les prédictions de la manière ou l'ordre dans lequel les données étaient divisées. Le résultat a montré une sensibilité de 99,11 %, une précision de 98,54 % et une spécificité de 98,25 %. Comme le montrent les résultats, il a été prouvé que l'approche NB pondérée surpassait l'approche NB régulière ainsi autres modèles. En raison de la méthodologie de recherche de grille utilisée dans le modèle, NB présentait certains inconvénients, comme le coût de calcul et les auteurs ont suggéré que ces lacunes pourraient être étudiées plus avant et améliorées à l'aide d'algorithmes génétiques.

Dans l'article [25],les auteurs ont présenté une étude sur la détection précoce du cancer du sein en utilisant quatre algorithmes d'apprentissage automatique appliqués aux données extraites de l'analyse sanguine. L'étude visait à comparer quatre algorithmes et à analyser les résultats des modèles d'apprentissage automatique. Les méthodes utilisées étaient k-NN, ANN, SVM et Extreme Learning Machine. Le réglage et l'optimisation des hyperparamètres ont également été utilisés pour améliorer les résultats de la classification. Le principal avantage fourni par cette optimisation d'hyperparamètres était qu'elle affectait la précision du système en fonction du nombre de neurones de la couche cachée et que la plage de ces plages pouvait être réglée manuellement par l'utilisateur. Le plus haut niveau d'exactitude a été démontré par le modèle ELM qui a donné une précision moyenne de 80 %. Le nombre optimal de couches de neurones cachés s'est avéré être de 1800 à la suite des différents tests effectués. L'utilisation d'ELM standard est plus avantageux en termes de précision. L'étude

[26] a présenté un modèle hybride d'apprentissage automatique pour le diagnostic du cancer du sein basé sur un système expert flou multicouche. Le modèle ELM-RBF utilise un classificateur connu sous le nom d'ELM ou Extreme Learning Machine, combiné à un noyau de fonction de base radiale. Ces auteurs ont également utilisé WBCD en se basant sur plusieurs métriques d'évaluation pour analyser le modèle. Ces métriques comprenaient les méthodes MAPE, RMSE, matrice de confusion, R2 et k-fold cross-validation (k = 10 dans ce cas). Après des tests approfondis, une précision moyenne de 98,05 % a été obtenue par ce modèle hybride ELM-RBF à l'aide d'une technique de validation croisée de 10 fois. Cependant, dans le même temps, un modèle SVM linéaire n'a montré qu'une précision de 90,56%. Les résultats obtenus par la mise en œuvre et l'analyse du modèle hybride ELM-RBF ont prouvé que ce modèle hybride surpassait le SVM linéaire dans presque tous les niveaux, et la précision obtenue avec le modèle hybride était également nettement supérieure à celle du SVM linéaire.

Dans [27], les auteurs ont analysé la classification de la détection du cancer du sein et les capacités de détection à l'aide du traitement d'images. Dans leur étude sur le traitement d'image, l'image a été prétraitée avant d'être utilisée pour éliminer les redondances présentes dans les images d'entrée sans affecter les images du produit final. La méthode proposée utilise les coefficients de transformée en ondelettes discrètes (DWT) comme vecteur de caractéristiques et utilise les algorithmes SVM et ANN. Des images de patients souffrant de problèmes de santé ont été traitées avec l'ondelette, le puissant outil mathématique d'extraction de caractéristiques, pour calculer les coefficients DWT. Les ondelettes sont particulièrement utiles pour la classification car elles fournissent des informations sur la fréquence d'un signal en fournissant des informations localisées. La précision rapportée était de 99,51 % pour SVM et de 98,54 % pour ANN. L'une des limites des RNA est la difficulté d'obtenir des résultats précis. Ainsi, plusieurs images pour détecter le cancer du sein en peu de temps ne peuvent pas être analysées avec cette méthode. Au début, le pré-traitement est effectué après la modification des dimensions de la résolution et du contraste des images d'entraînement et d'entrée a aidé à organiser toutes les images dans les mêmes proportions sans affecter les détails. En se basant sur la méthode DWT, différentes parties de l'image peuvent être identifiées pour la sélection des caractéristiques, une image grise est produite après l'extraction des caractéristiques, avec des paramètres de pixel allant de 0 à 255. Pour identifier les différences, nous avons utilisé SVM pour séparer les échantillons de cancer du sein en deux groupes, ceux touchés par le cancer et les tissus mammaires sains. Une technique d'extraction de caractéristiques et de traitement d'image est utilisée qui aide les radiologues à détecter le cancer du sein dans une phase préliminaire ou un stade plus précoce avec une plus grande précision.

L’étude [28] fournit deux nouvelles techniques évolutives, E(T)-DBN-BP-ELM et E(T)-DBN-ELM-BP, en combinant DBN avec un classificateur ELM (Extreme Learning Machine). En raison du vaste espace de solution des topologies DBN, l'algorithme génétique (GA), qui peut effectuer une recherche approfondie dans l'espace de solution de manière extraordinaire, a été utilisé pour l'optimisation architecturale dans les méthodes recommandées. La technique E(TW)-DBN a été utilisé dans ce travail en exploitant les algorithmes génétiques (GA) pour résoudre les deux difficultés, permettant à la topologie et aux poids DBN d'évoluer simultanément. Les auteurs de l’article [29] proposent de nombreux algorithmes d'exploration de données pour détecter le cancer du sein sur

l'ensemble de données WDBC. Dans la première étape, des algorithmes génétiques sont utilisés pour extraire des traits significatifs et pertinents. Dans la deuxième étape, diverses méthodologies d'exploration de données (systèmes de classification multiples) sont utilisées pour établir une décision pour deux groupes distincts d'individus avec et sans cancer du sein. Le modèle Rotation Forest avec 14 caractéristiques basées sur GA a fourni la plus grande exactitude de classification (99,48 %).

Cet article [30], compare six algorithmes d'apprentissage automatique sur le jeu de données WDBC : Classification and Regression Tree (CART), Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Linear Regression (LR), et Perceptron multicouche (MLP) en analysant l'exactitude des tests de classification, l'exactitude des données normalisées et la durée d'exécution. La collecte de données comprend 32 caractéristiques, que les méthodes statistiques réduisent à 12 (mode). Les chercheurs utilisent également le classificateur d'empilement (Voting Classifier) pour tester l'exactitude des apprenants de base comme la régression logistique (LR), l'arbre de décision (DT), le clustering de vecteurs de support (SVC), K-Nearest Neighbors (KNN), Random Forest (RF) et Naive Bayes (NB) (niveau méta). Cette étude [31] propose une nouvelle technique d'identification du cancer du sein basée sur des algorithmes d'apprentissage automatique. Les auteurs ont mené une enquête expérimentale sur un ensemble de données pour évaluer les performances. Par rapport aux méthodologies existantes, la méthode proposée a produit des résultats extrêmement précis et efficaces. Dans l'article [32], une nouvelle méthode intelligente de diagnostic du cancer du sein a été présentée en utilisant l'algorithme génétique de recuit simulé dirigé par le gain d'information (IGSAGAW) pour choisir les caractéristiques. Au cours de cette procédure, les praticiens classent les caractéristiques à l'aide de l'algorithme Information Gain (IG) avant d'extraire les meilleures caractéristiques optimales à l'aide de la méthode d'apprentissage par machine à vecteurs de support sensible aux coûts (CSSVM). En plus de réduire la complexité de la méthode SAGASW en extrayant efficacement le sous-ensemble de caractéristiques optimal, la technique de sélection de caractéristiques proposée atteint la précision de classification maximale et le coût de mauvaise classification le plus bas. La performance de KNN est régie par la valeur K, qui est le nombre d'éléments voisins.

L'étude de l'article [33] examine les performances de KNN en testant diverses fonctions de distance et valeurs de K pour découvrir un KNN efficace. Chaque expérience utilisant un ensemble de données consistait en trois tours. La première version de l'expérience ne contenait pas de sélection de fonctionnalités. La deuxième itération a utilisé la sélection de modèle basée sur la norme L1, qui a appliqué la sélection de caractéristiques de classificateur de vecteur de support linéaire, tandis que la troisième itération a utilisé la sélection de caractéristiques basée sur le Chi-square. En utilisant les fonctions de distance de Canberra ou de Manhattan, la méthode de sélection d'entités basée sur le Chi-square a atteint la plus grande exactitude pour les deux ensembles de données. Dans un autre sens les auteurs du travail [34] ont proposés un système d'aide à la décision clinique (CDSS) utilise une technique de co-évolution coopérative qui traite la sélection des fonctionnalités (FS) et la sélection des instances (IS) comme des sous-problèmes discrets. Dans ce travail, les caractéristiques et les instances sont sélectionnées à l'aide de la technique du Wrapper, qui combine la co-évolution coopérative et un classificateur de forêt aléatoire. L'ensemble de données réduit a été utilisé pour entraîner un classificateur de forêt aléatoire, qui a soutenu la

prise de décision clinique. Des techniques de visualisation de données et d'apprentissage automatique telles que la régression logistique, les k plus proches voisins, la machine à vecteurs de support, les bayésien naïf, les arbres de décision, la forêt aléatoire et la forêt de rotation ont été utilisées sur cet ensemble de données. L'objectif de l'étude [35] était d'effectuer une analyse comparative des applications de détection et de diagnostic du cancer du sein en utilisant la visualisation de données et l'apprentissage automatique. Le modèle de régression logistique avec toutes les caractéristiques a donné la précision de classification la plus élevée (98,1 %).

Dans l'article [36], les chercheurs proposent une nouvelle stratégie de sélection des caractéristiques basée sur les arbres de décision de renforcement des colonies d'abeilles et du gradient, dans le but de résoudre des problèmes tels que l'efficacité et la qualité informative des caractéristiques sélectionnées. Ils obtiennent une optimisation globale des entrées de l'arbre de décision en utilisant la méthode des colonies d'abeilles pour identifier les traits significatifs. La méthode génère l'espace des entités couvertes par le jeu de données. À l'aide d'un algorithme de colonie d'abeilles artificielles, les caractéristiques non pertinentes sont supprimées en fonction des informations décisionnelles qu'elles fournissent. Des expériences ont également été menées à l'aide de deux ensembles de données sur le cancer du sein. L'étude [37], a fourni un apprentissage d'ensemble de fonctionnalités basé sur Sparse Autoencoders et Softmax Regression pour catégoriser le cancer du sein comme bénin (non cancéreux) ou malin (cancéreux) en utilisant l'ensemble de données WDBC. De plus, la méthode proposée surpasse à la fois le modèle Stacked Sparse Autoencoders (SSAE-SM) et le modèle basé sur la régression Softmax. L'objectif de ce projet est d'étudier le pronostic automatisé du cancer du sein à l'aide des techniques d'apprentissage automatique et d'exploration de données.

Dans le papier [38], les auteurs ont proposé une méthodologie d'ensemble en couches qui utilise l'empilement et le vote comme approches de combinaison de classificateurs dans nos méthodes d'ensemble pour différencier les tumeurs bénignes des tumeurs malignes du sein. Chaque classificateur d'ensemble en couches a des "classificateurs" et des "MetaClassifiers ". Les MetaClassifiers sont capables d'inclure de nombreuses méthodes de classification. Dans ce travail, les auteurs ont construit des classificateurs d'ensembles imbriqués à deux couches. Les classificateurs d'ensemble à deux couches de MetaClassifiers utilisent deux ou trois algorithmes de classification indépendants. L'ensemble de données WDBC a été utilisé pour les tests et le modèle a été testé à l'aide de la méthode de validation croisée K-fold. Dans cette étude [39], les auteurs ont développé un modèle de prédiction en combinant une technique d'apprentissage basée sur l'intelligence artificielle avec une méthode statistique multivariée. L'exploration de données joue un rôle important dans l'automatisation du processus de diagnostic. Les ensembles de données accessibles à partir de divers référentiels sont bruités. Cette recherche fournit une technique de sélection de caractéristiques hybrides qui peut être utilisée avec l'ACP (Analyse en Composantes Principales) et un réseau de neurones artificiels (ANN). ACP effectue le prétraitement des données et l'extraction des caractéristiques.

Combinant l'algorithme slap swarm algorithm (SSA) avec l'optimisation par essaim de particules, l'article [40] propose une solution d'optimisation hybride pour le problème « Feature selection ». La combinaison des deux approches donne un algorithme nommé SSAPSO, qui augmente l'efficacité des procédures d'exploration et d'exploitation. Le SSAPSO est utilisé pour sélectionner la meilleure collection de fonctionnalités à partir de

plusieurs ensembles de données UCI, où les caractéristiques redondantes ou déroutantes sont éliminées de l'ensemble de données d'origine tout en préservant ou en augmentant la précision. L'article [41] propose une stratégie hybride qui combine deux algorithmes, l'optimisation du loup gris (GWO) et l'optimisation de l'essaim de particules (PSO), de telle manière que les fonctions critiques sont identifiées tandis que les fonctions sans conséquence et la complexité sont éliminées. Cela permet au gadget d'apprendre les tâches de catégorisation tout en entraînant le classificateur avec l'ensemble de données. Une stratégie hybride est principalement basée sur des algorithmes métaheuristiques d'intelligence d'essaim, qui imitent le comportement de gestion et de chasse du loup gris dans la nature, et PSO, dans lequel les individus se déplacent en réponse à leurs emplacements optimaux locaux et mondiaux. Dans le cadre de l'expérience, dix-sept ensembles de données de la bibliothèque d'apprentissage automatique de l'UCI ont été utilisés.

L'analyse de la littérature effectuée par les auteurs [42] a été basé sur la technique d'imagerie numérique infrarouge, qui suppose qu'une simple comparaison thermique entre un sein sain et un sein atteint de cancer révèle toujours une augmentation de l'activité thermique dans les tissus précancéreux et les régions entourant un cancer du sein en croissance. De plus, nous avons déterminé qu'un diagnostic assisté par ordinateur (CAD) employant un traitement d'image infrarouge était impossible sans un modèle similaire au modèle bien connu de l'hémisphère. Combinant des méthodes avancées de vision par ordinateur et des modèles d'apprentissage en profondeur, cette étude procède à une analyse comparative de plusieurs systèmes de détection du cancer du sein.

L'étude diagnostique du papier [43] utilise une stratégie d'optimisation métaheuristique inspirée du comportement de chasse au filet à bulles des baleines à bosse pour choisir et pondérer les attributs les plus efficaces extraits d'images microscopiques de cytologie mammaire et optimiser un classificateur de machine à vecteurs de support. À l'aide d'un ensemble de données sur le cancer du sein provenant du référentiel UCI, l'approche proposée a été évaluée. Plusieurs approches de validation, y compris des tests d'hypothèses statistiques (t-test et ANOVA), ont été utilisées pour valider les résultats de la catégorisation.

Les auteurs de l'article [44] ont suggéré une stratégie de diagnostic basée sur l'apprentissage automatique dans laquelle une machine à vecteurs de support était utilisée pour différencier les cas malignes des cas bénignes. Pour améliorer les performances de classification de la technique, ils ont appliqué les algorithmes de redondance minimale, de pertinence maximale et de Chi-square algorithme pour choisir des caractéristiques plus importantes de l'ensemble de données sur le cancer du sein. Les résultats des tests ont montré que le classificateur de la machine à vecteurs de support fonctionne mieux avec le sous-ensemble de fonctionnalités choisi à l'aide de la stratégie de pertinence maximale de redondance minimale. Dans cette étude [45], les auteurs ont évalué les plus récents modèles de détection et de classification du cancer du sein basés sur l'apprentissage automatique à l'aide d'une analyse comparative. Selon l'analyse comparative, You Only Look Once (YOLO) et RetinaNet sont les modèles les plus récents avec la plus grande précision basée sur des architectures de détection et de classification de base. Cette étude [46] examine la précision prédictive de nombreux algorithmes d'exploration de données disponibles pour la récurrence du cancer du sein. Dans un effort pour améliorer la précision du modèle de prédiction, il intègre l'optimisation des essaims de particules en tant que sélection de caractéristiques dans trois classificateurs bien

connus, y compris Bayesien naïf, le voisin le plus proche K et l'apprenant d'arbre de décision rapide.

Dans l'article [47], un modèle de classification robuste basé sur ANN a été créé pour augmenter la précision de la catégorisation du cancer du sein. La méthode Taguchi a été utilisée pour déterminer le nombre idéal de neurones pour la seule couche cachée de l'ANN. La sélection d'un nombre adéquat de neurones contribue à la résolution du problème de surajustement en affectant les performances de classification d'un ANN. Cela a abouti au développement d'un modèle de classification robuste pour la catégorisation du cancer du sein sur l'ensemble de données WDBC.

Le problème du déséquilibre des classes a été mis en évidence comme un obstacle important à la performance de classification d'un certain nombre d'approches d'apprentissage courantes. La méthode SMOTE peut être utilisée pour générer des points d'échantillonnage aléatoires afin d'améliorer le taux de déséquilibre. Cependant, son application est limitée par la génération de marginalisation et l'aveuglement de la sélection des paramètres.

Dans l'étude [48], les auteurs ont amélioré la méthode SMOTE de telle sorte que les nouveaux points d'échantillonnage sont distribués plus près du centre de l'échantillon minoritaire avec une plus grande probabilité, empêchant ainsi la marginalisation des données élargies. Les expériences révèlent que la méthode proposée surpasse l'algorithme SMOTE original tout en élargissant l'ensemble de données WDBC déséquilibrées. Dans cet article [49], les auteurs ont créé GeFeS, une technique de sélection de caractéristiques basée sur un Wrapper étendu qui utilise un algorithme évolutif parallèle intelligent unique (Algorithme génétique GA). Le GeFeS proposé empêche le surajustement et augmente considérablement la précision de la classification. Pour améliorer la précision, la robustesse et l'intelligence de l'AG, les chercheurs ont ajouté un nouvel opérateur pour la pondération des caractéristiques, mis à jour les opérateurs de mutation et de croisement et intégré la validation croisée imbriquée dans la procédure d'AG. Le classificateur k-plus proche voisin (KNN) a été utilisé pour évaluer l'utilité des attributs fournis. Ils ont évalué les performances de GeFeS sur des ensembles de données comme WDBC.

En utilisant l'analyse en composantes principales (ACP) et les modèles d'apprentissage automatique (MLM) basés sur les propriétés des données, les auteurs de [50] discutent d'un système de détection du cancer du sein qui permet aux utilisateurs d'identifier le cancer du sein. Pour identifier quel modèle a fourni la plus grande précision sur les données, cinq modèles différents ont été formés et évalués sur l'ensemble de données (diagnostic) sur le cancer du sein du Wisconsin. Pour une validation croisée de 10 fois, l'arbre du modèle logistique (LMT) a atteint une exactitude, une précision, une sensibilité et une mesure F de 97,53 %, 97,59 %, 95,75 % et 96,66 %, respectivement. Les auteurs de ce travail [51] ont examiné cinq algorithmes d'apprentissage automatique supervisé et les performances de l'étude ont été testées à l'aide de divers facteurs de mesure des performances statistiques. Cette recherche [52] propose un modèle hybride pour une détection efficace du cancer du sein qui combine plusieurs méthodes d'apprentissage automatique. Cette recherche présente une évaluation préliminaire de l'utilisation de l'apprentissage automatique pour prédire le pronostic du cancer du sein. Les auteurs de [53] ont examiné 1021 patientes opérées d'un cancer du sein dans leur institut, dont 610 d'entre elles. La récidive du cancer (à la fois loco-régionale et systémique) et la mortalité due à la maladie dans les 32 mois ont été

sélectionnées comme critères de jugement. Pour chaque résultat, ils ont créé deux types de modèles d'apprentissage automatique (ANN et SVM).

Cette recherche [54] fournit un examen comparatif des stratégies d'apprentissage automatique, d'apprentissage en profondeur (DL) et d'exploration de données utilisées pour la prédiction du cancer du sein. Leur objectif principal était de comparer plusieurs méthodes actuelles d'apprentissage automatique et d'exploration de données afin de déterminer la meilleure façon de prendre en charge des ensembles de données volumineux avec une précision de prédiction élevée. L'objectif principal de ce papier de revue était de mettre en évidence toutes les études précédentes sur les algorithmes d'apprentissage automatique utilisés pour la prédiction du cancer du sein, et cet article fournit toutes les informations nécessaires aux débutants qui souhaitent analyser les algorithmes d'apprentissage automatique afin d'acquérir une solide compréhension. L'étude [55] vise à donner un aperçu des nouvelles applications de l'apprentissage automatique et des technologies de l'apprentissage profond pour détecter et catégoriser le cancer du sein, ainsi qu'un aperçu du développement dans ce domaine. Cet article examine la catégorisation du cancer du sein à l'aide de l'imagerie médicale multimodalité. Il présente d'abord une introduction aux différentes méthodes d'apprentissage automatique, suivie d'une explication des différents algorithmes DL et architectures spécialisées pour la détection et la classification du cancer du sein. Pour offrir une compréhension globale du domaine, les auteurs présentent également une introduction rapide des différentes modalités d'image. Cette étude a été menée dans le même contexte, en utilisant un large éventail de bases de données de recherche comme source d'information pour l'accès à diverses publications de terrain.

Les auteurs [56] ont présenté une stratégie en deux étapes dans cet article : la régression logistique a été utilisée dans la première phase pour exclure les caractéristiques les moins pertinentes. Le réseau neuronal Group Method Data Handling (GMDH) a été utilisé dans la deuxième étape pour différencier les échantillons bénins et malins. Un modèle hybride basé sur des techniques d'optimisation et d'apprentissage automatique avec pondération des caractéristiques a été utilisé pour diagnostiquer le cancer du sein dans cet article [57]. Les auteurs ont utilisé la pondération des caractéristiques (FW) basée sur K-Means pour différencier davantage les échantillons bénins et malins.

Les auteurs de l'étude [58] ont étudié une approche d'apprentissage d'ensemble basée sur SVM pour la détection de BC. Cette étude se concentre sur la détection du BC et utilise une approche d'apprentissage d'ensemble basée sur SVM pour minimiser la variance diagnostique et améliorer la précision. Sur la base de la zone pondérée suggérée sous la technique WAUCE (Receiver Operating Characteristic Curve Ensemble), douze SVM distincts ont été hybrides. Pour construire des règles de classification fiables et interprétables à partir d'un ensemble d'arbres de décision pour la détection du cancer du sein, une approche améliorée d'extraction de règles basée sur la forêt aléatoire (RF) a été développée [59]. En premier lieu, un grand nombre de modèles d'arbres de décision ont été construits à l'aide de Random Forest pour produire une abondance de règles de décision. En deuxième lieu les règles de décision ont été extraites des arbres entraînés à l'aide d'une technique d'extraction de règles. Enfin, un algorithme évolutif multi-objectif amélioré (MOEA) a été utilisé pour trouver le prédicteur de règle optimal en termes de précision et d'interprétabilité.

Dans [60], des forêts à décision aléatoire ont été suggérées. La classification du cancer du sein dans cette publication peut être obtenue en combinant les avantages des classificateurs Feature Weight et Hyperparameter Tuned Random Decision Forest. En ce qui concerne les poids des caractéristiques, le clustering Kernel Neutrosophic C-Means a été utilisé, ce qui attribue des poids plus élevés aux caractéristiques pertinentes et des poids plus faibles aux caractéristiques moins applicables. Après cela, le modèle de classificateur Random Decision Forest a été réglé à l'aide de la technique d'optimisation bayésienne pour acquérir les meilleurs paramètres d'hyper-réglage. En utilisant la technique du Wrapper, l'étude a fourni une stratégie de sélection de caractéristiques pour identifier le sous-ensemble pertinent de caractéristiques pour la tâche d'apprentissage automatique [61]. La pondération des caractéristiques a été utilisée dans l'étude [62] pour créer une méthode de diagnostic assistée par ordinateur efficace pour le cancer du sein. En particulier, une technique d'encapsulation basée sur l'algorithme d'optimisation d'Ant Lion a été fournie, qui recherche simultanément les pondérations optimales des caractéristiques et les valeurs paramétriques d'un réseau neuronal multicouche. En ce qui concerne les paramètres du réseau de neurones, la sélection de neurones cachés et des techniques d'entraînement à la rétropropagation ont été utilisées.

Table II.1 - Etude comparative sur les recherches consacrées à la détection du cancer du sein à l'aide de différents algorithmes d'apprentissage automatique

Article	**Dataset**	**Algorithmes**	**Exactitude (%)**
[13]	WBCD	SVM, Naive Bayes, K-Nearest Neighbours, Decision Tree.	97.31
[14]	-	K-NN, SVM, Bayesian Network.	97.49
[15]	BCW	K Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), Stochastic Gradient Descent (SGD), Linear Support Vector Classifier (SVC (linear)), Extra Tree (TE), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Neural Network (NN).	96.49
[16]	BCW	ANN	81.37
[17]	WBCD	Support Vector Machine (SVM), Random Forest (RF), Bayesian Networks (BN), k-Nearest Neighbor (kNN).	-
[18]	WBCD	Multilayer Perceptron, kNN, Classification and Regression Trees, NB, SVM.	96.70
[19]	WBCD	Support Vector Machine, Naive Bayes, kNN, C4.5 decision trees.	97.13

[20]	WBCD	Bayesian network, J48.	97.80
[22]	-	K-Nearest Neighbours, NB, kNN.	91.6
[23]	WBCD	RVM, Naive Bayes, neural networks, DT, SVM.	97.0
[24]	WBC	Naïve Bayesian NB (weighted NB)	98.54
[25]	-	k-NN, ANN, SVM, Extreme Learning Machine.	80.0
[26]	WBCD	Fuzzy ELM-RBF.	98.05
[27]	-	SVM, ANN.	99.51
[30]	WDBC	Statistical Feature Selection Technique and stack ensemble of multiple ML classifier	95.17
[32]	WDBC et WBC	Information gain directed simulated annealing genetic algorithm wrapper	95.8
[33]	WDBC, WBC	Chi2 feature selection, KNN	98.62
[34]	WDBC	GA, Random Forest	97.1
[36]	WDBC	Artificial Bee Colony and Gradient Boosting Decision Tree	97.9
[37]	WDBC	Stacked sparse autoencoders and softmax regression	98.60
[38]	WDBC	SV-Na¨ıveBayes-3-MetaClassifiers	98.07
[39]	WBC	PCA, ANN	97.0
[40]	WDBC	Hybrid of SSA and PSO	97.8
[43]	WDBC	Whale optimization algorithm, PSO, GA and SVM	98.82
[63]	WBC	Expectation Maximization, Classification et Regression Trees, fuzzy rule-based reasoning method	94.1
[46]	WBC	PSO, Naïve	81.3
[47]	WDBC	Méthode Tugachi et ANN amélioré	98.80
[48]	WDBC	SMOTE algorithm	98.79
[49]	WDBC	Parallel GA with new operator and KNN	98.51

[50]	WBC	PCA, machine learning models (MLMs), logistic model tree (LMT).	97.53
[64]	WDBC	ANN	99.47
[56]	WDBC	Logistic Regression feature selection, Group method data handling neural network classifier	99.6
[57]	WDBC	Modified Harris Hawks Optimization, Extreme Learning Machine Embedded with Feature Weighting	98.76
[58]	WDBC	weighted area under the receiver operating characteristic curve model, Combined twelve different support vector machines	97.68
[59]	WDBC	Random Forest	95.09
[60]	WDBC	Random Decision Forests (RDF)	-
[61]	WDBC	Wrapper approach basée sur Binary Bat algorithm	93.54
[65]	WDBC	Grey Wolf Optimizer algorithm	94.82
[62]	WDBC	Feature weighting, Neural Networks	98.37
[66]	WDBC	Modified Bat Algorithm	-
[67]	WDBC	GA, SVM	98.77
[68]	216 images privées.	Particle Swarm Optimized Wavelet Neural Network	93.67
[69]	WDBC et autres jeux de données	PCA, SVM	97.89
[70]	WDBC et autres jeux de données	Krill Herd Algorithm	97.76
[71]	M. G Cancer Hospital & Research Institute, Visakhapatnam, India	Deep neural network with Support Value (DNNS).	97.21
[72]	Wisconsin	Support Vector Machine (SVM).	94.3
[73]	Department of Obstetrics and Gynaecology of the University of	Artificial neural network (ANN).	86.95

	Coimbra (CHEA)		
[74]	Wisconsin	Sequential Minimal Optimisation (SMO).	96.99
[75]	Wisconsin	k-Nearest Neighbours classifier (k-NN).	95.90
[76]	MRI from radiologists of the University of Bari Aldo Moro	Genetic algorithm (GA) optimized artificial neural network (ANN).	89.77
[77]	Thermograms from Federal Fluminense University Hospita	k-Nearest Neighbours classifier (k-NN).	94.44
[78]	DDSM	Gaussian Mixture Model (GMM).	86.0
[79]	Wisconsin	Decision Tree J48.	94.56
[80]	WDBC	fuzzy decision tree	94.53
[81]	WDBC	PCA avec algorithme K-NN	95.61

II.2.2. Prédiction des maladies cardiovasculaires à l'aide de l'apprentissage automatique

L'apprentissage automatique est une branche de l'intelligence artificielle de plus en plus utilisée dans le domaine de la médecine cardiovasculaire. C'est essentiellement la façon dont les ordinateurs donnent un sens aux données et décident ou classent une tâche avec ou sans supervision humaine. Le cadre conceptuel de l'apprentissage automatique est basé sur des modèles qui reçoivent des données d'entrée (par exemple, des images ou du texte) et, grâce à une combinaison d'optimisation mathématique et d'analyse statistique, prédisent les résultats (par exemple, favorables, défavorables ou neutres). Plusieurs algorithmes d'apprentissage automatique ont été appliqués aux activités quotidiennes. Par exemple, un algorithme d'apprentissage automatique commun appelé SVM peut reconnaître des modèles non linéaires à utiliser dans la reconnaissance faciale, l'interprétation de l'écriture manuscrite ou la détection de transactions frauduleuses par carte de crédit. Les algorithmes dits de boost utilisés pour la prédiction et la classification ont été appliqués à l'identification et au traitement des Spams. Un autre algorithme, noté forêt aléatoire, peut faciliter les décisions en faisant la moyenne de plusieurs nœuds. Alors que le traitement des réseaux de neurones convolutifs combine plusieurs couches pour la classification et la segmentation des images [82]–[84]. Nous avons précédemment décrit les détails techniques de chacun de ces algorithmes [85]–[87], mais aucun consensus n'a émergé pour guider la sélection d'algorithmes spécifiques pour une application clinique dans le domaine de la médecine cardiovasculaire. Bien qu'il soit possible de sélectionner des algorithmes optimaux pour les questions de recherche et de reproduire des algorithmes dans différents ensembles de données cliniques, l'interprétation clinique et le jugement pour la mise en œuvre d'algorithmes sont

très difficiles. Une compréhension approfondie des connaissances statistiques et cliniques des praticiens de l'apprentissage automatique est également un défi.

De nombreux travaux ont été réalisés pour la prédiction des maladies cardiaques à l'aide d'outils et de techniques d'apprentissage automatique, d'apprentissage en profondeur et d'exploration de données. Différents ensembles de données, algorithmes et méthodes utilisés par les chercheurs et les résultats observés ainsi que les travaux futurs sont menés pour trouver des méthodes efficaces de diagnostic médical des maladies cardiovasculaires. Partout dans le monde, la conception de modèles de prédiction du diagnostic des maladies cardiovasculaires est une recherche mobile depuis quelques décennies. La prédiction et le diagnostic automatiques des maladies cardiovasculaires constituent un problème médical dominant dans le monde réel. La détection des maladies cardiaques à ses débuts est cruciale pour un meilleur traitement. Différentes approches ont été utilisées par les chercheurs du monde entier pour la prédiction précoce des maladies cardiaques. La Table II.2 donne un aperçu des articles qui utilisent des approches d'apprentissage automatique appliquées pour la prédiction des maladies cardiovasculaire. Certaines des principales approches de prédiction des maladies cardiovasculaires sont les suivantes :

Les auteurs de l'étude [88] ont comparé différents systèmes d'apprentissage automatique pour la prédiction des maladies cardiovasculaire dans ses premières étapes. La qualité de l'ensemble de données a été améliorée en adoptant des techniques de prétraitement, l'accent étant mis principalement sur l'observation des données et le traitement des valeurs incorrectes et manquantes. De plus, ils ont utilisé les trois algorithmes d'apprentissage automatique supervisé pour prédire la condition, et leurs résultats ont été comparés à l'utilisation de diverses mesures d'évaluation. L'ensemble de données est collecté à partir du référentiel UCI. Deux jeux de données sont obtenus, un avec 303 occurrences et 14 caractéristiques et l'autre avec 1026 occurrences et 14 caractéristiques. Lorsque les jeux de données sont combinés, le jeu de données final compte 1 329 instances et 14 entités. À propos des résultats expérimentaux, l'ensemble de données avec une précision de 100 % pour l'entraînement et de 97,29 % pour l'ensemble de test, sont les plus élevés de tous les classificateurs. La fiabilité de l'évaluation SVM, DT et RF à l'aide d'une technique de validation croisée 10 fois. Avec une exactitude de 99,39 % dans le cas de RF en tête des autres algorithmes, suivi du DT avec une exactitude de 97,59 %.

Dans la recherche [89], la maximisation conditionnelle de l'information mutuelle (CMIM) a été utilisée pour identifier un sous-ensemble de polymorphisme nucléotidique unique (SNP) qui sont ensuite utilisés avec divers algorithmes d'apprentissage automatique, à savoir KNN, LDA, SVM, NB, ANN avec l'ensemble approche pour l'analyse. Parmi lesdits algorithmes, l'approche d'ensemble utilisant les classificateurs SVM, ANN et NB a donné la précision la plus élevée de 93,21 % et le score F1 de 91,27 %. La source de l'ensemble de données est la base Welcome Trust Case Control Consortium (WTCCC) contenant 2001 échantillons.

Dans le travail [90], les auteurs cherchent à identifier le meilleur algorithme pour prédire la maladie en utilisant plusieurs techniques de l'apprentissage automatique du référentiel UCI d'ensembles de données sur les maladies cardiaques. Dans la prédiction, l'ensemble de données se divise en deux ensembles, 80 % de données d'entraînement et 20 % de données de test, pour analyser la précision du modèle de classification. Les données obtenues ont été

analysées à l'aide des algorithmes KNN, RF, DT et SVM, et que RF est le meilleur modèle de classification avec une exactitude de 90 %.

Les auteurs de l'article [91] développent un modèle efficace pour la prédiction des maladies coronariennes, en utilisant des classifications d'apprentissage automatique telles que DT, RF, KNN, AdaBoost, Gradient Boosting, ainsi que des classificateurs hybrides. Relief et LASSO sont différentes approches d'évaluation utilisées pour déterminer les attributs les plus critiques à partir de références médicales en fonction des valeurs de classement. Les cinq ensembles de données sont Cleveland, VA Long Beach, la Suisse, la Hongrie et Statlog du référentiel UCI pour collecter et créer un ensemble de données plus vaste et plus fiable. Le chercheur traite également des problèmes de sur-ajustement et de sous-ajustement de l'apprentissage automatique. De plus, plusieurs approches hybrides, telles que Boosting et Bagging, permettent d'augmenter le taux de test tout en réduisant le temps de performance. Le modèle Random Forest with Bagging (RFBM) est précis à 99,05 %.

Dans le papier [92], les auteurs de cette étude ont proposé une approche de prédiction efficace et rentable pour la détection précoce des MCV. Un hôpital tertiaire du sud de l'Inde a fourni un total de 1670 dossiers de santé classés. 70 % des données obtenues servent à entraîner le modèle prédictif. Le système de prédiction utilise Python et cinq algorithmes d'apprentissage automatique de pointe NB, KNN, RF, AdaBoost et LR. La meilleure méthode de prédiction est la RF, qui a identifié avec succès 470 points de données médicales sur 501, donnant un taux d'exactitude de 93,8 %.

Les auteurs ont proposé que le système analyse les symptômes signalés par les données d'entrée de l'utilisateur et prédit l'apparition de la maladie en sortie [93]. La prédiction des maladies s'effectue en utilisant six techniques d'algorithme : SVM, KNN, RF, LR, DT et ANN avec une couche cachée. Ce projet donne également un aperçu de l'EDA. De plus, la plateforme Web prédit le risque pour la santé de l'utilisateur et propose des recommandations pertinentes en fonction de son état de santé. L'ensemble de données du référentiel UCI a été utilisé et prend en compte 13 attributs, qui constituent la base de base des tests et fournissent des résultats précis. Même sans restructuration, ce système fonctionne admirablement. Cette recherche pourrait répondre aux inquiétudes concernant la simplicité d'explication du modèle. De plus, avec une forêt aléatoire avec un taux d'exactitude de 93,60 %, la prévision est entièrement précise et réalisable.

Les auteurs du travail [94] utilisent des techniques d'apprentissage automatique pour développer un modèle de prédiction précoce des maladies cardiaques, qui profitera aux médecins pour reconnaître la maladie . Les algorithmes DT, LR, KNN, RF, SVM, NB et Gradient Boosting sont des algorithmes de classification. Ils sont utilisés pour prédire l'incidence des maladies cardiovasculaires. Leurs performances sont comparées à des mesures d'évaluation telles que l'exactitude, la sensibilité et la précision afin de déterminer le meilleur modèle de classification pour prédire les cas de maladie cardiaque. Le référentiel UCI de l'ensemble de données de Cleveland utilise pour l'entraînement et les tests, qui prennent en compte l'âge, le sexe, l'inconfort thoracique, la pression artérielle au repos, l'ECG au repos, la fréquence cardiaque maximale et produit une sortie indiquant si la personne a ou non l'infection cardiaque. L'utilisation de méthodes de sélection de caractéristiques contribue à la précision du modèle. Le classificateur de forêt aléatoire a atteint une exactitude de 93,44 %.

Dans [95], les auteurs ont développé un modèle hybride qui produit des résultats efficaces en utilisant une technique d'apprentissage automatique, qui se reconnaît dans les applications. Les probabilités obtenues à partir d'une technique d'apprentissage automatique intègrent l'autre technique d'apprentissage automatique dans un modèle hybride. L'ensemble de données Cleveland du référentiel UCI est utilisé pour l'analyse, et il existe 303 cas avec 14 caractéristiques différentes. Prédiction des maladies cardiovasculaires à l'aide d'algorithmes d'apprentissage automatique tels que Random Forest et Decision Tree. L'utilisation du modèle hybride avec Decision Tree et Random Forest pour améliorer la tâche. La découverte suggère qu'en utilisant la technique Random Forest et un modèle hybride, la détection des maladies cardiaques est efficace. L'arbre de décision a un taux d'exactitude de 79 %, tandis que la forêt aléatoire a un taux d'exactitude de 81 %, tandis que le modèle hybride a un taux de 88 %.

Les auteurs du papier [96] ont proposé un système d'apprentissage automatique qui utilise plusieurs méthodes pour prédire les risques de développer une maladie cardiaque. Cinq algorithmes sont utilisés pour exécuter le Framework : NB, RF, Logistic Model Tree, DT et SVM. Les algorithmes sont entraînés et testés en exécutant l'ensemble de données Cleveland. L'ensemble de données est collecté et traité ; ensuite, les fonctionnalités les plus importantes sont prises en compte pour la sélection des fonctionnalités. À cette fin, 80 % des données (242 instances) sont utilisées pour entraîner l'ensemble de données, et les 20 % restants (61 cas) ont été utilisés pour les tests. Pour la recherche, RF fournit le taux de précision le plus élevé avec 95,08 %.

Les auteurs du papier [97] ont effectué une évaluation comparative des performances de divers algorithmes de classification d'apprentissage automatique tels que DT, NB, RF et LR. L'objectif de l'étude était de déterminer le meilleur système ML efficace pour détecter les maladies cardiaques à l'aide de l'ensemble de données UCI Cleveland, qui a été rendu accessible pour analyse sur le site Web de Kaggle. Une description détaillée des 14 caractéristiques employées dans l'étude proposée. Le travail suggéré prédit les maladies cardiaques en étudiant les quatre algorithmes de classification énumérés ci-dessus et, en fonction des performances, en prédisant efficacement si le patient souffre ou non d'une maladie cardiaque. Le médecin insère les valeurs d'entrée du rapport médical de la personne. Les données de sortie sont converties en un système pour déterminer la probabilité de développer des problèmes cardiaques. Les résultats montrent que la RF est la méthode la plus efficace pour prédire les maladies cardiovasculaires, avec une exactitude de 90,16 %.

Les auteurs [98] ont proposé une approche de forêt aléatoire hybride avec un modèle linéaire (HRFLM). C'est un hybride de la forêt aléatoire et du modèle linéaire. Les caractéristiques essentielles des deux méthodes ont intégré les ensembles de données hongrois, Cleveland et suisse du référentiel UCI. KNN, DT, NB, LR, SVM, ANN, Generalized Linear Model, Gradient Boosted Trees et Genetic Algorithm sont des techniques d'apprentissage automatique qui classent la gravité de la maladie. HRFLM aussi une autre approche révolutionnaire utilisée, a assez bien prédit les maladies cardiaques. L'exactitude obtenue du HRFLM était de 88,7 %.

Les auteurs de l'étude [99] ont suggéré un Framework de prédiction des maladies cardiaques en utilisant Python et basé sur la méthode de forêt aléatoire utilisée pour l'entraînement et les tests des techniques d'apprentissage automatique à l'aide des ensembles de données Cleveland

du référentiel UCI des maladies cardiaques. Ces jeux de données contiennent 303 instances et 76 entités. L'ensemble de données utilisé pour l'entraînement d'algorithmes était de 75 % et de 25 % pour les tests. Les résultats ont été visualisés dans le Visual Studio Code à l'aide d'une interface utilisateur graphique. Pour la classification, le classificateur de forêt aléatoire a été utilisé et l'exactitude obtenue était de 97,56 %.

Dans le papier [100], les auteurs ont mené une étude comparative sur les maladies cardiaques en utilisant RF, DT et NB pour créer une méthode de prédiction permettant d'examiner et de prédire le potentiel des maladies cardiaques. L'ensemble de données Statlog du référentiel UCI, qui comprenait 270 cas et 13 attributs, a été utilisé pour l'entraînement et les tests du modèle. L'ensemble de données se divise en un modèle d'entraînement avec 80 % et pour les tests avec 20 %. Les résultats des tests indiquent que la technique RF a surpassé le DT et le NB dans la prédiction des maladies cardiaques. Comparée à d'autres algorithmes de prédiction des maladies cardiaques, la méthode RF donne les meilleurs résultats avec une précision de 81,0 %.

Les auteurs de cette étude [101], ont développé un modèle de prédiction en temps réel des maladies cardiovasculaires. La technologie se développe qui peut déterminer la présence et l'absence de maladie cardiaque chez un utilisateur, et elle est capable de prédiction et de diagnostic. L'ensemble de données du référentiel UCI des ensembles de données communs sur les maladies cardiaques de Cleveland, avec 303 cas et 76 caractéristiques, a été utilisé pour entraîner et tester l'algorithme. La petite taille de l'échantillon est divisée en 50 % pour l'entraînement et 50 % pour tester les données afin de créer les techniques d'apprentissage automatique NB, SVM, RF, LR, KNN et DT sont les méthodes d'apprentissage automatique utilisées. Les chercheurs révèlent le classificateur RF avec un taux de précision de 89,0 % pour la prédiction des maladies cardiovasculaires. Leur concept a été construit sur l'idée que les téléphones portables seraient toujours connectés à Internet, ce qui n'est pas toujours le cas.

Dans [102], les auteurs ont proposé une méthodologie de prédiction des maladies cardiaques visant à découvrir les traits essentiels à l'aide d'algorithmes d'apprentissage automatique, ce qui améliore la précision des prédictions. Au lieu de collecter des données à partir d'un référentiel Internet, les chercheurs ont collecté des informations auprès des hôpitaux et des entreprises de santé du district de Sylhet au Bangladesh pour créer un bon questionnaire et l'ensemble de données le plus précieux lié à la prédiction des maladies cardiaques. L'ensemble de données comprenait 564 occurrences et 18 attributs, et le modèle utilisait des algorithmes d'apprentissage automatique tels que LR, DT, KNN, SVM et NB, entre autres. La précision des différentes méthodes dépend des instances dans l'ensemble de données peut varier. SVM a produit les meilleurs résultats dans le modèle proposé, avec une précision de 91,0 % pour les instances de seuil de l'ensemble de données.

Les études de [103] visent à examiner le diagnostic des maladies cardiaques à l'aide du référentiel d'apprentissage automatique UCI des ensembles de données de Cleveland, contenant 303 cas et 14 attributs utilisés pour cette recherche et cette évaluation. Les ensembles de données avant d'être utilisés pour les analyses sont prétraités pour supprimer toutes les données manquantes et bruyantes. L'étude a comparé six approches d'apprentissage automatique distinctes basées sur plusieurs mesures de performance : LR, NB, KNN, SVM, RF et DT. L'analyse révèle que SVM fournit les meilleurs résultats, 89,34 %.

[104] Les auteurs ont développé un modèle de prédiction des maladies cardiaques qui est à la fois efficace et précis et basé sur des algorithmes d'apprentissage automatique. Le système s'appuie sur des techniques de classification telles que KNN, DT, ANN, NB, LR, SVM et des stratégies de sélection de caractéristiques conventionnelles telles que la pertinence maximale, la redondance minimale et le relief. La sélection de caractéristiques est une approche mutuelle d'information conditionnelle rapide unique utilisée pour surmonter le problème de sélection de caractéristiques. Ces techniques de sélection de caractéristiques ont été utilisées pour améliorer la précision de la classification et réduire le temps d'exécution. Par rapport à d'autres algorithmes, l'exactitude de SVM est de 92,37 %. La technique proposée est simple à adopter dans le domaine de la santé pour détecter les problèmes cardiaques.

Les auteurs de cette étude [105] ont proposé une application Web basée sur l'apprentissage automatique pour prédire les maladies cardiaques. Il affichera l'algorithme calculant la probabilité d'existence d'une maladie cardiaque et son résultat sur la page Web. De plus, cela réduit le temps et le coût de la prédiction de la maladie. Il existe plusieurs techniques d'apprentissage automatique ; les algorithmes s'entraînent à l'aide de l'ensemble de données Cleveland des conditions du référentiel UCI. Pour l'entraînement, 75 % des entrées de l'ensemble de données ont été utilisées, les 25 % restants étant utilisés pour tester la précision de l'algorithme. Les quatre algorithmes d'apprentissage automatique utilisés étaient DT, LR, NB et SVM. Des ensembles de données individuels ont été formés et testés pour chacun des quatre algorithmes. Différents aspects devaient être utilisés pour identifier l'algorithme le plus efficace. Avec une précision de 82,89 %, la technique LR est la plus efficace des quatre. SVM était précis à 81,57 %, tandis que DT et NB ont été précis à 80,43 % et 81,49, respectivement.

Dans [106], les auteurs ont comparé certaines techniques d'apprentissage automatique largement utilisées pour prédire les MCV. Les classificateurs des techniques d'apprentissage automatique ANN, SVM, DT et RIPPER utilisent WEKA 3.6 et le train modèle et sont testés à l'aide des ensembles de données Cleveland du référentiel UCI, contenant 303 cas et 14 fonctionnalités. Dans le prétraitement des données, la taille de l'échantillon de classificateur avec 296 occurrences. Les performances de l'algorithme sélectionné se comparent à d'autres classificateurs tels que ANN, NB et KNN. Le résultat du modèle a révélé que les techniques sélectionnées fonctionnaient mieux, SVM atteignant un taux d'exactitude de 90,0 %.

Les auteurs de [107] ont proposé un modèle de prédiction des maladies cardiovasculaires à l'aide de techniques d'apprentissage automatique. Divers systèmes neuronaux et exploration de données sont utilisés pour déterminer la gravité de la maladie cardiaque chez les patients. Le fait de ne pas identifier le problème tôt aurait un impact sur le cœur ou entraînerait un mort subite. Les données proviennent de la base de données Framingham du Kagglerepository et les algorithmes d'apprentissage automatique utilisés sont LR, RF, SVM, DT, J48, KNN, NB et AdaBoost. Le SVM s'avère être le meilleur, avec une précision de 90,3 %.

Les auteurs de cette étude [108] ont mené un travail de comparaison sur la prédiction des maladies cardiaques en utilisant des techniques d'apprentissage automatique telles que la classification KNN, DT et SVM. Ils ont utilisé l'ensemble de données VA Long Beach du référentiel UCI des maladies cardiaques pour l'entraînement et les tests d'algorithmes, contenant 270 instances et 12 fonctionnalités. Une matrice de confusion basée sur l'évaluation

menée pour l'exactitude, la spécificité et la sensibilité. Les résultats de leurs tests ont montré que SVM dépassait DT et KNN pour prédire les maladies cardiovasculaires, avec une spécificité de 83 %, une sensibilité de 100 % et une exactitude de 92,0 %.

Les auteurs de [109] ont proposé une conception préliminaire pour un système de prédiction des maladies cardiaques basé sur le cloud qui utiliserait des algorithmes d'apprentissage automatique. Deux ensembles de données du référentiel UCI des maladies cardiaques, l'ensemble de données Cleveland (303 instances avec 14 attributs) et l'ensemble de données VA Long Beach (270 cas avec 14 caractéristiques), ont été combinés pour entraîner des données détaillées. Dans les techniques d'apprentissage automatique, cinq opérations de classification et de prédiction sont utilisées dans WEKA, notamment SVM, RF, NB, Multi-Layer Perception et LR. SVM était le meilleur classificateur parmi les autres méthodes, avec 97,53 % d'exactitude de classification.

Dans [110], les auteurs comparent la classification SVM et ANN pour la prédiction des maladies cardiovasculaires basée sur la valeur prédictive positive. Leurs données d'échantillon provenaient de trois hôpitaux associés à l'Université des sciences médicales, en Iran, contenant 1324 cas et 25 attributs. Les échantillons ont été prélevés sur des patients hospitalisés entre 2016 et 2017 souffrants de maladies cardiovasculaires. Les données collectées à partir du référentiel UCI sont basées sur les variables mentionnées dans la directive sur la politique de données de Cleveland. Le prétraitement, la fusion, le filtrage, la normalisation et la compression des données ont été utilisées pour régir les données obtenues et téléchargées dans Microsoft Excel et SPSS, et le calcul statistique a été effectué à l'aide de R. Dans l'algorithme, l'ensemble de données a été divisé en 70 % pour l'entraînement et 30 % pour les tests. Leurs études ont révélé que l'algorithme ANN surpassait les performances en termes de sensibilité et de puissance et le taux d'exactitude était 91,75 %.

Les auteurs de l'article [111] ont développé un système de diagnostic basé sur MLP avec rétro propagation comme technique de formation. Les performances du système proposé examinent l'utilisation de la précision, de l'exactitude, de la spécificité et de la sensibilité et utilisent l'ensemble de données Cleveland sur les maladies cardiaques du référentiel UCI, avec 76 caractéristiques et 303 cas. Ces caractéristiques pour le prétraitement des ensembles de données ont exclu six cas avec des valeurs manquantes et ont choisi seulement 14 des 76 traits comme maladie cardiaque la plus critique. Le résultat expérimental a révélé que le modèle MLP-NN avait une exactitude élevée de 93,39 %.

Dans le papier [112], les auteurs ont créé de nombreux classificateurs d'apprentissage automatique et les appliquent pour créer le meilleur modèle de prédiction des maladies cardiaques. Le jeu de données utilisé dans cette étude provient du jeu de données Cleveland du référentiel UCI, qui comprenait initialement 303 occurrences et 76 caractéristiques. Plusieurs approches d'apprentissage automatique bien connues, notamment DT, ANN, SVM et NB, ont été étudiées pour développer, comprendre et interpréter divers modèles de prédiction des maladies cardiovasculaires. Comparé aux autres modèles, le modèle ANN avait l'exactitude la plus élevée avec 84,25 %.

Les auteurs de cette étude [113] ont utilisé une analyse comparative basée sur des techniques d'exploration de données pour créer un modèle de prédiction des maladies cardiovasculaires. Les données utilisées pour les tests provenaient de la base de données d'échocardiographie trans-thoracique(ETT), avec 336 cas et 24 caractéristiques. DT-J48, NN et NB sont trois

modèles d'apprentissage automatique importants utilisés pour analyser et effectuer des opérations de classification. Les résultats de leurs tests ont révélé que la classification NN était significativement plus performante dans la prédiction des maladies cardiovasculaires, avec un taux d'exactitude de 97,91 %.

Dans [114], les auteurs ont proposé une méthodologie pour prédire les maladies cardiaques qui démontre comment les données synthétiques peuvent être utilisées pour répondre aux fuites de données et éviter les restrictions des ensembles de données de recherche médicale limités. Les ensembles de données de substitution utilisent des observations synthétiques pour modéliser le système et comparer les résultats des tests de précision des prédictions DT, RF et LR. L'ensemble de données contient 303 cas avec 76 attributs tirés de l'ensemble de données de Cleveland de la bibliothèque UCI et prétraités en 279 instances avec 14 caractéristiques. Ils ont découvert que l'utilisation d'une validation croisée de 10 fois à l'aide d'algorithmes RF, DT et LR avec des données de substitution pouvait améliorer la stabilité des prédictions de 81 %. Ils ont amélioré l'exactitude de la prédiction des maladies cardiovasculaires à l'aide d'ANN avec des données de substitution d'environ 16 % à 96,7 %.

Les auteurs de [115] proposent des techniques d'apprentissage en profondeur et d'apprentissage automatique pour trouver les résultats et l'analyse du référentiel UCI de l'ensemble de données sur les maladies cardiaques sont Cleveland, Long Beach VA, Switzerland et Hungary. L'ensemble de données se compose de 14 caractéristiques et 303 cas sont utilisés pour analyser les performances des algorithmes avec une matrice de confusion. La forêt d'isolement gère les ensembles de données avec des caractéristiques non pertinentes et les ensembles de données qui sont des caractéristiques pertinentes pour de meilleurs résultats. Les algorithmes d'apprentissage automatique considèrent DT, SVM, LR, l'application de l'approche d'apprentissage en profondeur dans le modèle, une précision de 94,2 %.

Dans le papier [116], les auteurs ont montré de nombreuses caractéristiques des maladies cardiaques et un modèle basé sur des méthodes d'apprentissage automatique telles que les algorithmes RF, DT, KNN et NB. Il utilise des ensembles de données du Cleveland du référentiel UCI des patients atteints de maladies cardiaques et ils ont sélectionné 303 cas avec 14 caractéristiques dans la collection pour tester la précision de divers algorithmes. Les résultats montrent que KNN obtient le score de précision le plus excellent de 90,78 %. Les données sont prétraitées avant d'être utilisées dans le modèle. Les algorithmes qui ont donné satisfaction avec des résultats plus significatifs sont : KNN, NB et RF.

Le travail [117] montre une évaluation des performances des modèles d'algorithmes d'apprentissage automatique utilisés pour prédire les problèmes cardiaques. Plusieurs algorithmes d'apprentissage automatique sont utilisés pour diagnostiquer les maladies cardiaques chez les patients à un stade précoce. Les algorithmes DT, KNN, GB, SVM et LR sont utilisés pour cette tâche, et l'ensemble de données utilise le référentiel UCI des maladies cardiaques. Les environnements d'enseignement et de test sont créés dans le langage de programmation Python à l'aide du notebook Jupyter. L'entraînement et les tests sur les algorithmes d'apprentissage automatique ont révélé que le KNN avait une précision élevée de 85,7 %.

Les auteurs de l'article [118] ont comparé des modèles alternatifs d'ensembles de données sur les maladies cardiaques pour prédire efficacement les cas de maladie cardiaque avec des

caractéristiques limitées. Pour 1025 patients de Cleveland, Suisse, Hongrois et Long Beach V, l'ensemble de données couvre 14 attributs. Sur l'ensemble de données sur les maladies cardiaques, une méthode d'extraction de caractéristiques est utilisée pour supprimer les valeurs manquantes. Pour démontrer les performances des algorithmes de classification sélectionnés pour classer au mieux et/ou prédire les situations cardiaques, les méthodes ont utilisé NB, KNN, DT, SVM, JRip, Adaboost, Stochastic Gradient Decent et DT-J48. Après avoir utilisé plusieurs stratégies pour classer l'ensemble de données sur les maladies cardiaques, le classificateur KNN a atteint une précision de classification de 99,70 %.

Les auteurs de cette étude [119], ont effectué des recherches approfondies dans MATLAB sur des modèles d'apprentissage supervisé tels que KNN, SVM, ANN et la rétro propagation multicouche à anticipation. L'ensemble de données Cleveland sur les maladies cardiaques du référentiel d'apprentissage automatique de l'UCI a été utilisé, qui comprenait 303 instances avec 76 attributs. L'ensemble de données effectue un prétraitement pour éliminer les entrées avec des valeurs manquantes. La taille des données résultantes était de 270 cas avec 13 attributs. 50 % des informations sont utilisées pour entraîner les modèles, tandis que les 50 % restants servent à les tester. Selon les résultats expérimentaux, SVM a surpassé la précision de la classification, avec un taux de réussite de 85,0 %.

Dans [120], les auteurs ont fourni un modèle utilisant la classification neuronale ML pour l'infection par les maladies cardiovasculaires. Le cadre du système neuronal reconnaît 13 caractéristiques et prévoit la présence ou l'absence de troubles cardiaques et de nombreuses mesures d'exécution chez le patient. L'ensemble de données du modèle a été extrait des ensembles de données de Cleveland du référentiel UCI d'apprentissage automatique et a examiné 76 attributs et 303 cas. Les algorithmes utilisés sont NB, Stochastic Gradient Descent, KNN, DT, SVM et Adaboost pour démontrer les performances de classification sélectionnées afin de classer les meilleurs et de prévoir les cas de maladies cardiaques. L'arbre de décision a obtenu 99,70 % de précision.

Les auteurs de cette étude [121] ont proposé un système qui vise à appliquer des méthodes d'ensemble pour améliorer la précision de la prédiction des maladies cardiaques. Des techniques d'apprentissage automatique sont utilisées pour prédire la phase initiale de la maladie cardiaque. L'ensemble de données est obtenu à partir du référentiel Kaggle. Les approches d'ensemble, c'est-à-dire les méthodes d'ensachage et de renforcement, utilisent deux techniques d'extraction de caractéristiques (analyse discriminante linéaire et analyse en composantes principales) pour identifier les caractéristiques essentielles de l'ensemble de données. Une comparaison des méthodes d'ensemble (ensachage et renforcement) et cinq classificateurs tels que SVM, KNN, DT, NB et RF fonctionnent sur des fonctionnalités sélectionnées. La méthode d'ensachage utilisant l'arbre de décision et l'approche d'extraction des caractéristiques d'analyse en composantes principales a atteint la plus grande précision de 98,6 % sur la base d'observations expérimentales.

Les auteurs de l'étude [122] ont présenté une étude comparative sur la prédiction des maladies cardiaques et ont tiré des conclusions analytiques. Selon les résultats de l'étude, les approches algorithmiques DT, KNN, ANN, NB, RF et LR améliorent la précision du système de prédiction des maladies cardiaques dans divers scénarios. Ils ont choisi comme emplacement pour cette base de données de Cleveland. Une matrice de corrélation utilisée pour analyser les données dans l'analyse et un diagnostic plus avancé pour des études

avancées. Sur la base du résultat, l'arbre de décision est le meilleur en raison de sa grande précision de 98,02 %, de sa haute précision et de sa faible MSE. En comparaison, LR et KNN ont une précision inférieure de 89,0 %.

L'étude [123] a mené une analyse comparative en appliquant des techniques d'apprentissage automatique pour prévoir les maladies cardiaques. Les algorithmes évalués comprenaient DT, NB, LR, SVM et RF. Le référentiel UCI des maladies cardiaques a fourni 303 cas et 14 caractéristiques pour les ensembles de données standard de Cleveland. La technique de validation croisée 10 fois utilisée lors de l'apprentissage et du développement du modèle. Selon les résultats de l'étude, l'algorithme de l'arbre de décision avait la meilleure précision de 93,19 % pour prédire la maladie cardiaque, suivie du SVM à 92,30 %.

Dans le papier [124] les auteurs ont développé un modèle hybride pour prédire les maladies cardiaques qui fusionne les techniques d'apprentissage automatique dans un cadre unique. L'ensemble de données a été extrait du référentiel Cleveland UCI des maladies cardiaques, y compris 303 cas et 14 attributs dans le modèle pour entraîner et tester les algorithmes. Le prétraitement des données minimise le nombre de caractéristiques de 14 à 12. Les algorithmes incluent la classification KNN, NB, RF, GA, J48 et SVM, avec leurs précisions, spécificités et sensibilités respectives dans la prédiction des maladies cardiovasculaires prises en compte. Les expériences ont révélé que SVM et NB avaient de meilleurs résultats pour prédire les maladies cardiaques avec une précision similaire de 89,2 %.

Un modèle prédictif de détection des maladies cardiaques cardiovasculaires basé sur divers indicateurs liés au cœur a été créé [125]. Les auteurs ont utilisé un algorithme d'apprentissage automatique pour créer ce modèle. Trois algorithmes supervisés distincts, à savoir RF, NB et J48 Classifier ont été utilisés pour prétraiter l'ensemble de données. La découverte de connaissances de la base de données (KDD) est utilisée pour extraire l'ensemble de données sur les maladies cardiaques du laboratoire ERIC, qui se composait de 209 cas de test. L'exactitude, la précision, le rappel et la mesure F, entre autres mesures de référence, ont été utilisés pour évaluer les performances de l'algorithme. Avec une exactitude de classification de 100 %, le modèle le plus efficace pour prédire les patients atteints de maladie cardiaque était la RF mise en œuvre sur des attributs spécifiés.

Les auteurs [126] ont créé une technique d'apprentissage automatique pour la prédiction des maladies cardiovasculaires basée sur la régression logistique (LR). Ils ont utilisé le module SK-Learn du logiciel Python pour comparer l'approche LR avec d'autres techniques telles que NB, KNN, SVM, DT et J48. Trois hôpitaux collaborant avec l'Université des sciences médicales AJA en Iran ont fourni des données sur les problèmes cardiaques. Il y a 1 324 instances et 25 fonctionnalités dans l'échantillon. Les résultats expérimentaux ont montré que l'algorithme LR fonctionnait mieux avec une précision de 86,89 %.

Les auteurs de cette étude [127], ont créé un modèle hybride basé sur une architecture d'apprentissage automatique pour diagnostiquer les patients atteints de maladies cardiovasculaires à l'aide de sept algorithmes de classification de premier plan en Python à savoir NB, KNN, LR, ANN, DT, SVM et Multilayer Perception. L'ensemble de données de Cleveland (303 instances et 76 attributs) a été utilisé pour utiliser une stratégie de validation croisée de 10 fois pour l'entraînement et les tests de modèles. Les meilleures caractéristiques des symptômes de maladie cardiaque ont été choisies à l'aide de techniques de sélection de caractéristiques pour éliminer les cas avec des valeurs manquantes. Les données ont été

affinées et évaluées par tous les classificateurs avec chaque algorithme de sélection de caractéristiques pour trouver le modèle le plus performant. Les résultats des tests ont révélé que la LR avec une validation croisée de 10 fois avait la meilleure précision de 89,0 % lorsqu'elle était sélectionnée par l'algorithme (feature selection) Relief. C'est une meilleure méthode de prédiction en termes de précision en raison des excellentes performances de LR avec Relief.

Table II.2 - Principales approches de prédiction des maladies cardiovasculaires.

Article	**Dataset**	**Algorithmes**	**Exactitude (%)**
[88]	UCI repository	Support Vector Machine, Decision Tree, Random Forest.	99.39
[89]	WTCCC	KNN, LDA, SVM, NB, ANN.	93.21
[90]	UCI repository	Decision Tree, Random Forest, Support Vector Machines, K Nearest Neighbors.	90.0
[91]	UCI repository: Cleveland, VA Long Beach, Switzerland, Hungary, Statlog	DT, RF, KNN, AdaBoost, Gradient Boosting, hybrid classifiers.	99.05
[92]	-	NB, KNN, RF, AdaBoost, and LR	93.8
[93]	UCI repository	SVM, KNN, RF, LR, DT et ANN.	93,6
[94]	UCI repository	DT, LR, KNN, RF, SVM, NB et Gradient Boosting.	93.44
[95]	UCI repository : Cleveland	Random Forest, Decision Tree.	88.0
[96]	UCI repository : Cleveland	NB, RF, Logistic Model Tree, DT, SVM.	95.08
[97]	UCI repository : Cleveland	DT, NB, RF, LR	90.16
[98]	UCI repository: Cleveland, Switzerland, Hungary	KNN, DT, NB, LR, SVM, ANN, Generalized Linear Model, Gradient Boosted Trees, Genetic Algorithm	88.7
[99]	UCI repository : Cleveland	Random Forest	97.56
[100]	UCI repository: Statlog	RF, DT, NB.	81.0

[101]	UCI repository: Cleveland	NB, SVM, RF, LR, KNN, DT	89.0
[102]	-	LR, DT, KNN, SVM, NB	91.0
[103]	UCI repository: Cleveland	LR, NB, KNN, SVM, RF, DT.	89.34
[104]	-	KNN, DT, ANN, NB, LR, SVM.	92.37
[105]	UCI repository: Cleveland	DT, LR, NB, SVM.	81.57
[106]	UCI repository: Cleveland	ANN, SVM, DT.	90
[107]	Kagglerepository's Framingham database	LR, RF, SVM, DT, J48, KNN, NB, AdaBoost	90.3
[108]	UCI repository: VA Long Beach	KNN, DT, SVM.	92.0
[109]	UCI repository: Cleveland, VA Long Beach	SVM, RF, NB, Multi-Layer Perception, LR.	97.53
[110]	-	SVM, ANN	91.75
[111]	UCI repository: Cleveland	MLP	93.39
[112]	UCI repository: Cleveland	DT, ANN, SVM, NB	84.25
[113]	Transthoracic Echocardiography	DT-J48, NN, NB.	97.91
[114]	UCI repository: Cleveland	DT, RF, LR.	96.7
[115]	UCI repository: Cleveland, VA Long Beach, Switzerland, Hungary	DT, SVM, LR.	94.2
[116]	UCI repository: Cleveland	RF, DT, KNN, NB	90.78
[117]	-	DT, KNN, GB, SVM, LR	85.7
[128]	UCI repository	LR, DT, KNN, SVM.	87.0

[118]	UCI repository: Cleveland, VA Long Beach, Switzerland, Hungary	NB, KNN, DT, SVM, JRip, Adaboost, Stochastic Gradient Decent, DT-J48	99.7
[119]	UCI repository: Cleveland	KNN, SVM, ANN, multilayered feed-forward backpropagation.	85.0
[120]	UCI repository: Cleveland	NB, Stochastic Gradient Descent, KNN, DT, SVM et Adaboost	99.7
[121]	-	SVM, KNN, DT, NB, RF.	98.6
[122]	UCI repository: Cleveland	DT, KNN, ANN, NB, RF, LR.	89.0
[123]	-	DT, NB, LR, SVM, RF.	92.3
[124]	UCI repository: Cleveland	KNN, NB, RF, GA, J48, SVM.	89.2
[125]	-	RF, NB, J48.	100
[126]	-	NB, KNN, SVM, DT, J48.	86.89
[127]	-	NB, KNN, LR, ANN, DT, SVM, Multilayer Perception	89.0

II.2.3. Aperçu des approches d'apprentissage automatique utilisées pour prédire le COVID-19

L'épidémie de maladie (COVID-19) causée par le nouveau coronavirus, le SRAS-CoV-2, s'est propagée rapidement dans le monde depuis la fin de 2019, touchant plus de 200 pays et régions début mai. L'Organisation mondiale de la santé (OMS) a déclaré l'épidémie de COVID-19 comme urgence de santé publique de portée internationale (USPPI) le 31 janvier[7] et l'a classée comme pandémie le 11 mars[8]. À l'échelle mondiale, au 4 juillet 2020, il y avait eu 10 922 324 cas confirmés de COVID-19, dont 523 011 décès, avec une moyenne de plus de 100 000 nouveaux cas confirmés par jour (Fig. 1), signalés à l'OMS (OMS, 2020e). Plus de 2,72 millions de cas ont été confirmés aux États-Unis, avec plus de 128 000 décès. En outre, le Brésil, la Russie, l'Inde, la Grande-Bretagne, l'Espagne, le Pérou et le Chili ont diagnostiqué plus de 250 000 personnes au total, et plus de 15 pays ont diagnostiqué plus de 100 000 personnes au total. Le COVID-19 a suscité une vive inquiétude à l'échelle internationale quant à la propagation de l'épidémie et à sa tendance de développement. De nombreuses études mathématiques se sont concentrées sur la modélisation du développement du COVID-19 et l'effet des interventions sur le confinement de la propagation. Comme le

[7] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200131-sitrep-11-ncov.pdf
[8] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf

reconnaît l'OMS, les modèles mathématiques opportuns jouent un rôle clé dans la prise de décision fondée sur des données probantes par les décideurs en matière de santé.

Plus tard, des techniques plus avancées basées essentiellement sur l'intelligence artificielle sont employées. L'utilisation de l'intelligence artificielle pour lutter contre le COVID-19 a servi à la prévention et au suivi des patients infectieux. En effet, en utilisant les coordonnées géographiques des personnes, certains gouvernements ont pu limiter leurs déplacements et localiser les personnes avec lesquelles ils étaient en contact. Le deuxième aspect dont bénéficie l'intelligence artificielle est la capacité de classer les individus, qu'ils soient touchés ou non. Enfin, l'intelligence artificielle offre la possibilité de faire une prédiction sur d'éventuelles contaminations futures. À cette fin, l'apprentissage automatique, souvent confondu avec l'intelligence artificielle, est précisément utilisé. Au-delà des différents algorithmes de l'apprentissage automatique, Neural Network est l'un des plus utilisés pour résoudre des problèmes du monde réel ce qui donne l'émergence de l'apprentissage profond.

Dans ce contexte, cette section donne un aperçu des recherches en apprentissage automatique effectuées pour traiter les données COVID-19. Un résumé des approches d'apprentissage automatique pour la détection, le diagnostic et la prédiction des cas de COVID-19, est présenté dans la Table II.3.

Les auteurs de cette étude [129] ont appliqué le modèle Support Vector Machine (SVM) pour la détection et la classification des cas de COVID-19. Les informations cliniques et les données des analyses de sang et d'urine ont été utilisées dans leur travail pour valider les performances de SVM. Les résultats de la simulation ont démontré l'efficacité du modèle SVM en atteignant une exactitude de 81,48 %, une sensibilité de 83,33 % et une spécificité de 100 %.

Les auteurs du papier [130] ont proposé une nouvelle approche basée sur l'hybridation du SVM avec le seuillage à plusieurs niveaux pour détecter les patients infectés par le COVID-19 à partir d'images radiographiques. Les performances de l'approche hybride ont été évaluées à l'aide de 40 images radiographiques pulmonaires à contraste amélioré (15 normales et 25 avec COVID-19). Un travail similaire a été effectué par d'autres auteurs [131], dans lequel une approche combinée basée sur la combinaison de SVM avec 13 modèles CNN pré-entraînés pour la détection du COVID-19 à partir d'images radiographiques thoraciques a été proposée. Les résultats expérimentaux ont montré que ResNet50 combiné à SVM surpasse les autres modèles CNN combinés à SVM en atteignant une exactitude de classification moyenne de 95,33 %.

Dans le papier [132], les auteurs ont utilisé le modèle SVM pour prédire les patients COVID-19 présentant des symptômes graves/critiques. 220 enregistrements d'observations cliniques/de laboratoire et 336 cas de patients infectés par le COVID-19 divisés en ensembles de données d'entraînement et de test ont été utilisés pour valider les performances du modèle SVM. Les résultats de la simulation ont montré que le modèle SVM atteint une aire sous la courbe (AUC) de 0,9996 et 0,9757 dans l'ensemble de données d'apprentissage et de test, respectivement.

Quatre approches d'apprentissage automatique (SVM avec Bagging Ensemble, CNN, Extreme Learning Machine (ELM) et Online Sequential ELM (OS-ELM)) ont été utilisé dans le travail [133] pour la détection automatique des cas de COVID-19. La performance

des approches proposées a été testée à l'aide d'ensembles de données de 702 images de tomodensitométrie (344 avec COVID-19 et 358 normales). Les résultats expérimentaux ont révélé l'efficacité de SVM avec Bagging Ensemble en obtenant une exactitude, une précision, une sensibilité, une spécificité, un score F1 et une ASC de 95,70 %, 95,50 %, 96,30 %, 94,80 %, 95,90 % et 95,80 %, respectivement.

Dans le travail [134] les auteurs ont proposé les moindres carrés-SVM (LS-SVM) et la moyenne mobile intégrée autorégressive (ARIMA) pour la prédiction des cas de COVID-19. Un ensemble de données de cas confirmés de COVID-19 collectés dans cinq des pays les plus touchés a été utilisé pour valider les modèles proposés. Il a été démontré que le modèle LS-SVM surpasse le modèle ARIMA en obtenant une exactitude de 80 %. Les auteurs de cette étude [135] ont appliqué des approches d'apprentissage automatique telles que SVM, Decision tree (DT) et KNN pour la détection automatique des cas positifs de COVID-19. Les performances des approches proposées ont été validées sur une base de données publique de radiologie COVID-19 divisée en ensembles d'entraînement et de test avec des taux de 70 % et 30 %, respectivement. En conséquence, les résultats les plus efficaces ont été assurés par le classificateur SVM avec une exactitude de 98,97 %, une sensibilité de 89,39 %, une spécificité de 99,75 % et un score F de 96,72 %.

Les auteurs du papier [136] ont utilisé SVM avec Naive Bayes (NB), Gradient boosting decision tree (GBDT), AdaBoost, CNN et Multilayer perceptron (MLP) pour un diagnostic rapide du COVID-19. Un ensemble de données de 980 images de tomodensitométrie (430 avec COVID-19 et 550 normaux) a été utilisé dans la simulation et les résultats ont montré que SVM surpasse les autres approches d'apprentissage automatique en atteignant une exactitude, une précision, une sensibilité et un score F1 moyens de 99,20 %, 98,19 %, 100 % et 99,0 %, respectivement.

Un modèle de régression linéaire pour la prédiction des patients infectés par le COVID-19 a été élaboré dans [137]. Les images CT de 52 patients recueillies dans cinq hôpitaux à Ankang, Lishui, Zhenjiang, Lanzhou et Linxia ont été utilisées pour évaluer les performances du modèle de régression. Les résultats de la simulation ont démontré que le modèle de régression linéaire surpasse l'algorithme Random Forest. Les modèles LR et RF ont montré une sensibilité et une spécificité de 1,0 et 0,89, 0,75 et 1,0 dans le test respectivement. Un autre travail similaire a été effectué [138], dans lequel un modèle de régression logistique de l'opérateur de réduction et de sélection au minimum absolu (LASSO) a été proposé. L'efficacité du modèle proposé a été évaluée sur la base d'images CT prises chez 196 patients (151 patients non sévères et 45 patients sévères). Les résultats expérimentaux ont montré la haute performance du modèle proposé par rapport aux paramètres CT quantitatifs et au score PSI en atteignant une exactitude de 82,70 %, une sensibilité de 82,20 %, une spécificité de 82,80 % et une AUC de 89 %.

Les auteurs de [139] ont proposé un modèle de régression supervisé, appelé XGBoost, pour prédire les patients COVID-19. Une base de données d'échantillons de sang de 485 patients infectés dans la région de Wuhan, en Chine, a été utilisée dans des simulations, et les résultats ont montré que XGBoost donne de bonnes performances en atteignant une précision globale de 90 % dans la détection des patients atteints de COVID-19. Dans [140], les auteurs ont utilisé le modèle de régression linéaire avec SVM et ANN pour la prédiction des patients infectés par COVID-19. L'efficacité des modèles proposés a été évaluée sur la base de

l'ensemble de données épidémiologiques recueillies à partir de nombreux rapports sur la santé de cas en temps réel. Les résultats de la simulation ont démontré que SVM a l'erreur absolue moyenne la plus faible avec une valeur de 0,21, tandis que le modèle de régression à l'erreur quadratique moyenne la plus faible avec une valeur de 0,46.

Une technique de régression linéaire avec un modèle mathématique SEIR (Susceptible, Exposed, Infectious, Recovered) a été proposée dans [141] pour prédire l'épidémie de COVID-19. Il a été testé à l'aide de données collectées à partir du référentiel de l'Université John Hopkins en tenant compte de la métrique de l'erreur quadratique moyenne du journal (RMSLE). Les résultats de la simulation ont montré que le modèle SEIR a le RMSLE le plus bas avec la valeur de 1,52. Dans cette recherche [142], la régression logistique avec forêt aléatoire, la régression partielle des moindres carrés (PLSR), le filet élastique et l'analyse discriminante flexible en sac (BFDA) ont été proposées pour prédire la gravité des patients COVID-19. L'efficacité des modèles proposés a été évaluée à l'aide des données de 183 patients gravement infectés par le COVID-19 et les résultats ont montré que le modèle de régression logistique surpasse les autres modèles d'apprentissage automatique en atteignant une sensibilité de 89,20 %, une spécificité de 68,70 % et une ASC de 89,20 %. Un autre travail similaire a été effectué [143], dans lequel six approches d'apprentissage automatique telles que l'apprentissage d'ensemble d'empilement (SEL), la régression de vecteur de support (SVR), la régression cubiste (CUBIST), la moyenne mobile intégrée auto-régressive (ARIMA), la régression de crête (RIDGE) et la forêt aléatoire (RF) ont été utilisées à des fins de prédiction dans les ensembles de données COVID-19.

Dans [144] les auteurs ont utilisé trois approches d'apprentissage automatique (régression linéaire, régression polynomiale et SVR) pour la prédiction et l'analyse de l'épidémie de COVID-19. Un ensemble de données contenant le nombre total de cas positifs au COVID19 a été collecté dans différents pays tels que la Corée du Sud, la Chine, les États-Unis, l'Inde et l'Italie. Les résultats ont montré la supériorité de la RVS par rapport à la régression linéaire et à la régression polynomiale. La précision moyenne pour la SVR, la régression linéaire et la régression polynomiale est de 99,47 %, 65,01 % et 98,82 %, respectivement.

Les auteurs du travail [145] ont proposé quatre modèles de régression linéaire (régression binomiale pénalisée (PBR, arbres d'inférence conditionnels (CIR), linéaire généralisé (GL) et SVM avec noyau linéaire) pour le diagnostic de COVID-19. Images CT et données cliniques recueillies auprès de 106 patients ont été utilisés dans la simulation et les résultats ont montré que SVM avec noyau linéaire donne de meilleurs résultats par rapport aux autres modèles en fournissant une exactitude de 0,88, une sensibilité de 0,90, une spécificité de 0,87 et une AUC de 0,92.

Les auteurs de [146] ont proposé une régression logistique avec six approches d'apprentissage automatique (Adaboost, Stochastic Gradient Boosting, Decision Tree, SVM, Multinomial Naïve Bayes et Random Forest) pour la détection et la classification du COVID-19. Il a été évalué à l'aide de 212 rapports cliniques répartis en quatre classes, notamment COVID, ARDS, SARS et Both (COVID, ARDS). Les résultats de la simulation ont montré que la régression logistique offre d'excellentes performances en obtenant 94 % de précision, 96 % de sensibilité, une exactitude de 96,20 % et 95 % du score F1.

Dans le papier [147], un arbre de décision boosté par gradient (GBDT) a été proposée avec arbre de décision, régression logistique et forêt aléatoire pour le diagnostic de COVID-19. 27

tests de laboratoire de routine collectés auprès du New York Presbyterian Hospital/Weill Cornell Medicine (NYPH/WCM) ont été utilisés pour évaluer cette technique. Les résultats expérimentaux ont révélé l'efficacité du GBDT en atteignant une sensibilité, une spécificité et une AUC de 76,10 %, 80,80 % et 85,40 %, respectivement.

Dans le travail [148], les auteurs ont développé un nouveau modèle (PBRR) en combinant la régression bayésienne de la crête (BRR) avec un polynôme à n degrés pour prévoir la progression de l'épidémie de COVID-19. Les performances du modèle PBRR ont été validées à l'aide d'ensembles de données publiques collectées auprès de l'Université John Hopkins disponibles jusqu'au 11 mai 2020. Les résultats expérimentaux ont révélé les bonnes performances du PBRR avec une exactitude moyenne de 91%.

Une méthode de forêt aléatoire sensible à la taille de l'infection (iSARF) a été proposée [149] pour le diagnostic du COVID-19. Un ensemble de données de 1020 images CT (1658 avec COVID-19 et 1027 avec pneumonie) a été utilisé pour évaluer les performances de l'iSARF. Les résultats de la simulation ont démontré que l'iSARF offre de bonnes performances en produisant une sensibilité de 90,7 %, une spécificité de 83,30 % et une précision de 87,90 % sous une validation croisée quintuple. Dans une autre étude [150], les auteurs ont combiné le modèle RF avec l'algorithme AdaBoost pour la prédiction de la gravité de la maladie COVID-19. L'efficacité du modèle RF amplifié a été évaluée sur la base des données géographiques, de voyage, de santé et démographiques du patient COVID-19. Le modèle RF boosté donne une exactitude de 94 % et un score F1 de 86 % sur l'ensemble de données utilisé.

Dans le travail [151], sept approches d'apprentissage automatique (Random Forest, Logistic Regression, KNN, Decision Tree, Extremely Randomized Trees, Naïve Bayes et SVM) ont été proposées pour l'identification des patients COVID-19 positifs. Des examens sanguins de routine collectés auprès de 279 patients ont été utilisés dans la simulation et les résultats ont démontré la faisabilité et l'efficacité de l'algorithme Random Forest en atteignant une exactitude, une précision, une sensibilité, une spécificité et une ASC de 82 %, 83 %, 92 %, 65 %, et 84 %, respectivement.

Les auteurs de [152] ont développé une méthode hybride (ARIMA-WBF) basée sur l'hybridation du modèle ARIMA et du modèle de prévision basé sur les ondelettes (WBF) pour prédire le nombre de cas confirmés quotidiens de COVID-19. L'efficacité d'ARIMA-WBF a été validée à l'aide d'ensembles de données de 346 cas provenant de cinq pays (70 : Canada, 71 : France, 64 : Inde, 76 : Corée du Sud et 65 : Royaume-Uni). Les résultats de la simulation ont montré les performances et la robustesse d'ARIMA-WBF dans la prédiction des cas de COVID-19.

Dans ce papier [153], les auteurs ont proposé une technique de génération de caractéristiques, appelée Residual Exemplar Local Binary Pattern (ResExLBP) avec Iterative ReliefF (IRF) et cinq méthodes d'apprentissage automatique (arbre de décision, discriminant linéaire, SVM, kNN et discriminant de sous-espace) pour la détection automatique du COVID-19. L'efficacité du modèle proposé a été validée à l'aide d'ensembles de données d'images radiographiques collectées sur le site Web GitHub et le site Kaggle. Les résultats de la simulation ont montré que ResExLBP avec IRF et SVM offre de meilleures performances par rapport aux autres modèles en fournissant une exactitude de 99,69 %, une sensibilité de 98,85 % et une spécificité de 100 %.

Les auteurs de [154] ont utilisé le MLP avec KNN, SVM, les arbres de décision et la forêt aléatoire pour l'identification du COVID-19 dans les images radiographiques thoraciques. L'efficacité des modèles proposés a été évaluée sur la base de la base de données RYDLS-20 de 1144 images radiographiques thoraciques divisées en ensembles d'entraînement et de test avec des taux de 70 % et 30 %. Les résultats expérimentaux ont montré la supériorité du MLP par rapport aux autres approches d'apprentissage automatique en fournissant un F1-Score de 89%.

Albahri et al [155] ont utilisé un modèle d'apprentissage automatique combiné à une nouvelle méthode de décision multicritère (MCDM) pour l'identification des patients infectés par le COVID-19. L'efficacité du modèle proposé a été évaluée sur la base d'images d'échantillons de sang. Les résultats de la simulation ont révélé que le modèle proposé est un bon outil pour identifier les cas infectés de COVID-19. Dans cette étude [156], les auteurs ont développé un modèle hybride basé sur la technique FbProphet et le modèle logistique pour la prédiction des tendances épidémiques de COVID-19. Le modèle hybride a été validé à l'aide de données épidémiologiques de séries chronologiques COVID-19 et les résultats ont révélé l'efficacité du modèle hybride pour la prédiction du tournant et de la taille de l'épidémie de COVID-19. Dans une autre étude [157], les auteurs ont proposé un système de détection assistée par ordinateur (CAO) basé sur l'apprentissage automatique (COVIDiag) pour le diagnostic du COVID-19. La performance de COVIDiag a été évaluée à l'aide d'images CT de 612 patients (306 avec COVID-19 et 306 normaux). Les résultats expérimentaux ont démontré l'efficacité de COVIDiag par rapport à SVM, KNN, NB et DT en atteignant la sensibilité, la spécificité et la précision de 93,54 %, 90,32 % et 91,94 %, respectivement.

Wang et al [158] ont proposé un modèle CNN profond, appelé Residual Network34 (ResNet34), pour le diagnostic du COVID-19 dans les images de tomodensitométrie. L'efficacité de ResNet34 a été validée à l'aide d'images de tomodensitométrie recueillies auprès de 99 patients (55 patients atteints de pneumonie virale typique et 44 patients atteints de COVID-19). Les résultats de la simulation ont montré que ResNet34 atteint une précision globale de 73,10 %, une spécificité de 67 % et une sensibilité de 74 %.

Trois techniques préformées sont proposées dans [159], notamment ResNet50, InceptionV3 et InceptionResNetV2 pour le diagnostic et la détection automatiques du COVID-19. Les études de cas comprenaient quatre classes comprenant des patients normaux, COVID-19, bactériens et de pneumonie virale. Les auteurs ont démontré que ResNet50 donne la plus grande précision dans trois ensembles de données différents.

Maghdid et al [160] ont proposé un modèle CNN avec AlexNet pour le diagnostic de COVID-19. Un ensemble de données de 361 images CT et 170 images radiographiques de la maladie COVID-19 collectées à partir de cinq sources différentes a été utilisé dans la simulation. Les résultats quantitatifs ont démontré qu'AlexNet atteint une précision de 98 %, une sensibilité de 100 % et une spécificité de 96 % dans les images radiographiques, tandis que le modèle CNN modifié atteint 94,10 % de précision, 90 % de sensibilité et 100 % de spécificité des images CT. Dans un autre sens huit modèles d'apprentissage en profondeur (DL) ont été proposés dans [161] (réseau entièrement convolutif (FCN-8 s), UNet, VNet, 3D UNet++, réseau à double chemin (DPN-92), Inceptionv3) pour la détection de COVID -19. L'efficacité des modèles proposés a été évaluée à l'aide de 1136 images CT (723 avec

COVID-19 et 413 normales) collectées dans cinq hôpitaux. Les résultats de la simulation ont démontré la supériorité de 3D UNet++ par rapport aux autres modèles CNN.

Dans les images de tomodensitométrie, UNet++ a été utilisé par les auteurs [162] pour la détection de COVID-19. La performance de UNet++ a été évaluée sur la base d'un ensemble de données de 106 images CT scan. Les résultats de la simulation ont montré que UNet++ fournit une précision par patient de 95,24 %, une sensibilité de 100 % et une spécificité de 93,55 %. Une précision par image de 98,85 %, une sensibilité de 94,34 % et une spécificité de 99,16 % ont également été obtenues.

Apostolopoulos et al [163] ont proposé cinq modèles CNN profonds (VGG19, MobileNetv2, Inception, Xception et Inception ResNetv2) pour les cas de détection de COVID-19. Les modèles proposés ont été testés à l'aide de deux ensembles de données de 1428 et 1442 images, respectivement. Dans le premier ensemble de données (224 avec COVID-19, 700 avec une pneumonie bactérienne et 504 normaux), l'approche MobileNetv2 a fourni de meilleurs résultats avec une précision du problème à deux classes, une précision du problème à trois classes, une sensibilité et une spécificité de 97,40 %, 92,85 %, 99,10 % et 97,09 %, respectivement. Dans le deuxième ensemble de données (224 avec COVID-19, 714 avec une pneumonie bactérienne et 504 normaux), l'approche MobileNetv2 a également fourni de meilleures performances en atteignant une précision du problème à deux classes, une précision du problème à trois classes, une sensibilité et une spécificité de 96,78 %, 94,72 %, 98,66 % et 96,46 %, respectivement.

Un autre modèle CNN profond a été développé [164] qui est composé de trois composants (un réseau dorsal, une tête de classification et une tête de détection d'anomalies). Cette technique a été évaluée à l'aide de 100 images radiographiques thoraciques de 70 patients extraites du référentiel Github. 1431 images supplémentaires de radiographie pulmonaire de 1008 patients extraites des données publiques de radiographie pulmonaire14 ont également été utilisées pour faciliter l'apprentissage en profondeur. Les résultats de la simulation ont montré que le modèle proposé est un outil de diagnostic efficace pour le dépistage COVID-19 à faible coût et rapide en atteignant une précision de 96 % pour les cas COVID-19 et de 70,65 % pour les cas non COVID-19. Un autre projet intéressant a été réalisé [165], dans lequel un Bayesian Convolutional Neural Networks (BCNN) a été utilisé en conjonction avec des Dropweights pour le diagnostic et la classification du COVID-19.

Dans [166], les auteurs ont proposé un modèle CNN, appelé CAPSNET, pour un diagnostic rapide et précis des cas de COVID-19. Le modèle CAPSNET a été évalué à l'aide de deux ensembles de données de 2 100 et 13 150 cas, respectivement. Dans le premier ensemble de données (1050 avec COVID-19 et 1050 sans résultats), CAPSNET a fourni des meilleurs résultats en obtenant une exactitude, une précision, une sensibilité, une spécificité, un score F1 de 97,23 %, 97,08 %, 97,42 %, 97,04 % et 97,24. % respectivement. Dans le deuxième ensemble de données (1050 avec COVID-19, 1050 non trouvés et 1050 pneumonies), CAPSNET a fourni aussi des meilleures performances en atteignant une exactitude, une précision, une sensibilité, une spécificité et un score F1 de 84,22 %, 84,61 %, 84,22 %, 91,79 % et 84,21 % respectivement.

Hammoudi et al [167] ont étudié six modèles CNN profonds (ResNet34, ResNet50, DenseNet169, VGG19, InceptionResNetV2 et RNN-LSTM) pour le dépistage et la détection du COVID-19. Un jeu de données de 5 863 images radiographiques d'enfants (normaux et

pneumonies) a été exploité pour évaluer les techniques proposées. Les résultats de la simulation ont montré que DenseNet169 surpasse les autres modèles CNN profonds en obtenant une précision moyenne de 95,72 %.

Les auteurs de cette étude [168] ont proposé deux modèles CNN profonds (AlexNet et InceptionV4) pour le diagnostic et l'analyse du pronostic des cas de COVID-19. L'efficacité des modèles proposés a été évaluée à l'aide de 5800 images CT divisées en 80 % d'entraînement et 20 % de test. Il a été démontré qu'AlexNet surpasse InceptionV4 en atteignant une précision globale de 94,74 %, une sensibilité de 87,37 % et une spécificité de 87,45 %. Bai et al. [73] ont fait un travail similaire en proposant un modèle EfficientNet B4 CNN avec un réseau de neurones entièrement connecté pour la détection et la classification des cas de COVID-19. Les images CT scan de 521 patients ont été utilisées dans la simulation.

Dans le travail [169], les auteurs ont proposé trois approches CNN approfondies (Alexnet, Googlenet et Restnet18) avec le modèle GAN pour la détection du COVID-19. Les approches proposées ont été évaluées à l'aide de trois scénarios : i) quatre classes (normal, pneumonie virale, pneumonie bactérienne et images COVID-19) ; ii) trois classes (COVID-19, Normal et Pneumonie) ; et iii) deux classes (COVID-19, Normal). Les résultats expérimentaux ont démontré que Googlenet offre de meilleures performances dans le premier et troisième scénario en atteignant une précision de 80,60 % et 100 %, respectivement. Alexnet fournit de meilleurs résultats dans le deuxième scénario en atteignant une précision de 85,20 %.

Singh et al [170] ont proposé une nouvelle approche d'apprentissage en profondeur basée sur des réseaux de neurones convolutifs avec évolution différentielle multi-objectifs (MODE) pour la classification des patients COVID-19. De plus, Mukherjee et al. [76] ont proposé un modèle CNN peu profond et léger pour la détection automatique des cas de COVID-19 à partir des radiographies pulmonaires d'une manière similaire.

Dans [171] Les auteurs ont proposé une approche efficace basée sur la combinaison du modèle CNN avec la méthode de classement et la technique SVM pour la détection du COVID-19. Les études de cas comprenaient deux ensembles de données générés à partir de 150 images CT, chaque ensemble de données contient 3000 images normales et 3000 avec COVID-19. Les résultats de la simulation ont montré les hautes performances et la robustesse de l'approche proposée par rapport aux modèles VGG16, GoogleNet et ResNet50 en termes de précision, de sensibilité, de spécificité, de sensibilité, de score F1 et de métriques du coefficient de corrélation de Matthews (MCC).

[172] Les auteurs ont proposé deux modèles CNN (MobileNetV2, SqueezeNet) combinés avec SVM pour la détection de COVID-19. L'efficacité des modèles proposés a été validée à l'aide d'un ensemble de données d'images radiographiques divisées en trois classes : normal, avec COVID-19 et avec pneumonie. La précision obtenue dans leur travail est de 99,27%.

[173] Les auteurs ont proposé une technique d'apprentissage par transfert profond ResNet50 pour la détection et la classification des patients infectés par COVID-19. L'efficacité de ResNet50 a été évaluée à l'aide de 852 images CT collectées à partir de divers ensembles de données (413 COVID-19 (+) et 439 normaux ou pneumonies). Les résultats de la simulation ont montré que le modèle ResNet50 offre des performances efficaces en atteignant une

spécificité, une précision, une sensibilité et une exactitude de 94,78 %, 95,19 %, 91,48 % et 93,02 %, respectivement.

Les auteurs de cette étude [174] ont proposé un modèle CNN profond modifié (Modified InceptionV3) pour le dépistage du COVID-19 sur les radiographies pulmonaires. Le Modified InceptionV3 a été évalué à l'aide de deux ensembles de données de radiographie pulmonaire, le premier ensemble de données a été collecté à partir de Github, le second a été collecté à partir de la base de données de la plateforme d'imagerie QUIBIMcovid19 et de divers référentiels publics. Les résultats expérimentaux ont montré que le modèle InceptionV3 modifié donne une précision, une ASC, une sensibilité et une spécificité moyennes de 76 %, 93 %, 93 % et 91,80 %, respectivement.

Chowdhury et al [175] ont introduit huit CNN profonds (DenseNet201, RestNet18, MobileNetv2, InceptionV3, VGG19, ResNet101, CheXNet et SqueezNet) pour la détection du COVID-19. Un ensemble de données de 3487 images radiographiques (423 avec COVID-19, 1485 avec pneumonie virale et 1579 normales) avec et sans augmentation d'image a été utilisé dans la validation des modèles proposés. Les résultats de la simulation ont montré que CheXNet donne de meilleurs résultats lorsque l'augmentation d'image n'a pas été appliquée avec une exactitude, une précision, une sensibilité, une spécificité, un score F1 de 97,74 %, 96,61 %, 96,61 %, 98,31 % et 96,61 % respectivement. Cependant, lorsque l'augmentation d'image a été utilisée, DenseNet201 surpasse les autres modèles CNN profonds en atteignant une exactitude, une précision, une sensibilité, une spécificité et un score F1 de 97,94 %, 97,95 %, 97,94 %, 98,80 % et 97,94 %, respectivement.

Les auteurs de cette étude [176] ont proposé un modèle CNN profond modifié basé sur la combinaison de Xception et ReNet50V2 pour détecter le COVID-19 à partir d'images radiographiques thoraciques. Le modèle proposé a été testé à l'aide de 11 302 images radiographiques pulmonaires (31 avec COVID-19, 4 420 avec pneumonie et 6 851 cas normaux). Les résultats expérimentaux ont montré que le modèle combiné donne une exactitude, une précision, une sensibilité et une spécificité moyennes de 91,4 %, 72,8 %, 87,3 % et 94,2 %, respectivement. Dans un travail similaire, Abbas et al. [91] ont adapté un modèle de réseau neuronal convolutif, appelé Decompose Transfer Compose (DeTraC). L'efficacité du modèle DeTraC a été validée à l'aide d'un ensemble de données d'images radiographiques collectées dans plusieurs hôpitaux et institutions du monde entier. Comme résultats, une précision de 95,12 %, une sensibilité de 97,91 % et une spécificité de 1,87 % ont été obtenues.

Table II.3 - Résumé des approches d'apprentissage automatique pour la détection, le diagnostic et la prédiction des cas de COVID-19

Article	**Dataset**	**Algorithmes**	**Exactitude (%)**
[129]	Tongji Hospital Affiliated to Huazhong University of Science and Technology	Support Vector Machine	81.48
[130]	-	Support Vector Machine	97.48

[131]	-	SVM, CNN	95.33
[132]	-	Support Vector Machine	-
[133]	COVID-19 CT scan images	SVM, CNN, ELM, OS-ELM	95.70
[134]	-	LS-SVM, ARIMA	80.0
[135]	COVID-19 radiology database	SVM, DT, KNN	98.97
[136]	COVID-19 CT scan images.	Support Vector Machine (SVM), Naive Bayes (NB), Gradient boosting decision tree (GBDT), AdaBoost, CNN, Multilayer perceptron (MLP)	99.2
[137]	CT images	Random Forest, Logistic regression	-
[138]	CT images	LASSO	82.7
[139]	-	XGBoost	90.0
[140]	-	SVM, ANN	
[141]	John Hopkins University repository	SEIR (Susceptible, Exposed, Infectious, Recovered)	-
[142]	Optical Valley Branch of Tongji Hospital, Wuhan	Logistic Regression, Random Forest, Partial Least Squares Regression (PLSR), Elastic Net, Bagged Flexible Discriminant Analysis (BFDA)	-
[144]	-	Linear Regression, Polynomial Regression, SVR.	99.47
[145]	CT images	Penalized binomial regression (PBR, Conditional inference trees (CIR), Generalised linear (GL), and SVM with linear kernel	88.0
[146]	-	Logistic regression, Adaboost, Stochastic Gradient Boosting, Decision Tree, SVM, Multinomial Naïve Bayes, Random Forest	96.2
[147]	New York Presbyterian Hospital/Weill Cornell Medicine (NYPH/WCM)	Decision Tree, Logistic Regression, Random Forest	-
[148]	John Hopkins University	PBRR	91.0

[149]	CT images	iSARF	87.9
[150]	-	Random Forest, AdaBoost	94.0
[151]	-	Random Forest, Logistic Regression, KNN, Decision Tree, Extremely Randomized Trees, Naïve Bayes et SVM	82.0
[153]	X-ray images	Decision tree, linear discriminant, SVM, kNN, and subspace discriminant	99.69
[154]	X-ray images	MLP, KNN, SVM, Decision Trees, Random Forest	-
[158]	CT images	ResNet34	73.10
[159]	X-ray images	ResNet50, ResNet101, ResNet152, InceptionV3, Inception–ResNetV2	99.70
[160]	X-ray images	AlexNet	98.0
[161]	CT images	Eight DL models (FCN–8 s, UNet, VNet, 3D UNet++, DPN–92, Inceptionv3, ResNet50, Attention ResNet50	-
[162]	CT images	UNet++	95.24
[163]	X-ray images	VGG19, MobileNetv2, Inception, Xception, InceptionResNetv2	96.78
[164]	X-ray images	CNN	-
[165]	X-ray images	Bayesian CNN	89.82
[166]	X-ray images	CAPSNET	97.23
[167]	X-Ray images	ResNet34, ResNet50, DenseNet169, VGG-19, InceptionResNetV2, RNN-LSTM	95.72
[177]	CT images	AlexNet, VGG16, VGG19, SqueezeNet, GoogleNet, MobileNetV2, ResNet18, ResNet50, ResNet101, Xception	99.51
[168]	CT images	AlexNet and Inception-V4	94.74
[169]	X-ray images	Alexnet, Googlenet, and Restnet18	100
[170]	CT images	CNN	93.50
[171]	CT images	CNN, SVM	98.27

[172]	X-Ray images	CNN, SVM	99.27
[173]	CT images and clinical data	ResNet-50	93.02
[178]	X-Ray & CT images	baseline CNN, VGG16, VGG19, DenseNet201, InceptionResNetV2, InceptionV3, Xception, Resnet50, MobileNetV2	92.60
[179]	X-ray images	VGG16, InceptionV3, Xception, DenseNet201, NasNetmobile	99.26
[174]	CT images	X-ray images	76
[175]	X-Ray images	CheXNet, DenseNet201, RestNet18, MobileNetv2, InceptionV3, VGG19, ResNet101, and SqueezNet	97.74
[176]	X-ray images	CNN	91.4

II.3. Méthodes de prédiction basées sur l'ontologie

De nos jours, la technologie s'est améliorée dans le monde entier et est devenue une partie essentielle de notre vie. Il aide les médecins à analyser et à diagnostiquer les problèmes médicaux et les maladies. A l'aide de l'intelligence artificielle en médecine, la science est devenue très demandée aujourd'hui. L'utilisation de l'intelligence artificielle dans de nombreux secteurs se généralise, car elle contribue à améliorer les soins de santé à bien des égards. Cependant, le projet d'IA est vulnérable à certains types de problèmes de santé, tels que les données non structurées, le temps de retard, etc. Par conséquent, de nouvelles approches basées sur l'ontologie doivent être intégrées, par exemple, la prédiction des maladies à l'aide de techniques d'ontologie et d'apprentissage automatique.

Les ontologies peuvent soutenir le diagnostic des maladies, en particulier en raison de leur capacité inhérente à traiter l'interopérabilité sémantique. Par conséquent, l'approche ontologique fournit un cadre sémantique dans lequel les données des patients pour l'évaluation et la gestion de leurs maladies peuvent être saisies, et les risques, profils et recommandations dérivés. L'approche basée sur l'ontologie peut également soutenir l'évaluation de la qualité des données cliniques utilisées. Une approche basée sur l'ontologie pour identifier les patients atteints de maladies chroniques est une approche sémantique avec des contraintes définies, des concepts et des relations prédéfinis, y compris son propre vocabulaire. Ce modèle d'ontologie prend des requêtes, communique avec la base de connaissances et analyse les dossiers des patients par le biais d'annotations sémantiques et, dans ce cas, identifie les patients atteints de diabète en intégrant des langages sémantiques. L'approche basée sur l'ontologie utilise également diverses fonctions telles que la gestion terminologique, l'intégration et le partage de données, la réutilisation des connaissances et l'aide à la décision. Les ontologies doivent représenter la réalité tout en ayant une base théorique solide.

D'un autre côté, l'ontologie est l'une des approches les plus adoptées pour gérer, organiser et extraire des données au cours des décennies précédentes. C'est une méthode de représentation de données qui a été mise en œuvre avec succès dans une variété de domaines, en particulier le domaine médical. Il est important en informatique en raison de sa capacité à représenter divers concepts et leurs relations dans différentes disciplines. En réalité, aucune ontologie unique n'est suffisante pour répondre aux demandes croissantes des soins de santé d'aujourd'hui, et les ontologies doivent être intégrées à des algorithmes d'apprentissage automatique pour prendre en charge l'intégration et l'analyse des données.

Dans ce contexte, cette section donne un aperçu des recherches en relation avec l'intégration de l'ontologie avec l'apprentissage automatique pour la prédiction dans le domaine de la santé en particulier, et dans d'autres domaine en générale.

Dans cet article [180], les auteurs ont proposé une méthode pour identifier les patients dont le niveau de risque de maladie diabétique. Dans ce travail sur le diabète, le niveau de risque du patient a été détecté en utilisant des techniques d'ontologie et d'apprentissage automatique. L'ontologie contient les symptômes, les causes et les traitements de la maladie. Dans l'apprentissage automatique, l'algorithme naïve bayésien est utilisé pour prendre des décisions sur le dossier du patient et définit également les possibilités de niveau de risque. L'algorithme proposé sera évalué par rapport aux paramètres suivants, à savoir la matrice de confusion, le niveau de précision, la moyenne. Ce travail proposé s'avère avoir un meilleur niveau de prédiction par rapport aux travaux existants.

Divakar et al [181] ont proposé un modèle qui permettra de prédire l'état de santé de l'homme en fonction des activités qu'il effectue pour prévenir les maladies cardiovasculaires. Ainsi, dans le cas d'un centre de soins de santé pour représenter l'état actuel des soins de santé à l'aide de réseaux sociaux, ayant différentes méthodes conventionnelles en se basant sur l'ontologie. WordSet est la source de l'ontologie où l'information est présente. Ces informations sont présentées dans le Web sémantique profond, elles sont considérées comme des entrées pour déterminer l'état de santé cardiovasculaire en fonction de l'activité d'une personne partagée sur les réseaux sociaux en ligne permettant l'accès entre les personnes et les lieux.

Connaître les antécédents du patient est essentiel pour qu'un médecin puisse effectuer une évaluation appropriée des risques du patient et suggérer un traitement approprié, c'est l'objectif de la recherche [182], dans laquelle les auteurs ont proposé un système basé sur une ontologie pour collecter les antécédents du patient et évaluer les facteurs de risque du patient dus aux antécédents de tabagisme, d'alcoolisme, de dysfonction érectile et d'antécédents cardiovasculaires. Selon les antécédents du patient, un score total est calculé pour chacun des facteurs ci-dessus. Selon le score, l'ontologie effectue l'évaluation des risques sur un profil de patient et prédit les risques potentiels et les complications du patient. Ces systèmes recueillent de meilleurs antécédents médicaux que les infirmières/diététistes/autres membres du personnel hospitalier à l'aide d'une questionnaire. Le questionnaire n'est pas statique, il est de nature adaptative, donc seules les questions pertinentes selon le contexte du patient seront posées. Le système réduit le nombre de questions, économisant ainsi le temps des patients. Puisque les questions sont lues à partir de l'ontologie, si nous voulons ajouter ou mettre à jour ou supprimer des questions, nous devons le faire uniquement dans l'ontologie du questionnaire. Les informations sur chaque patient

sont générées dans un fichier OWL séparé. Il peut être visualisé à partir de Protégé au format OWL et également à partir du système au format tabulaire. Le risque patient associé à différents facteurs est également généré. Ces valeurs de risque (score) aideront le médecin à comprendre la situation actuelle d'un patient. Le raisonnement basé sur l'ontologie permet de découvrir de nouvelles connaissances.

Un nouveau système d'analyse des risques de maladies chroniques basé sur une ontologie est décrit, qui permet la création d'une représentation globale des connaissances (ontologie) et d'une modélisation personnalisée pour un système d'aide à la décision [183]. Un modèle informatisé axé sur l'organisation des connaissances liées à trois maladies chroniques et aux gènes a été développé dans une représentation ontologique capable d'identifier les interrelations pour l'évaluation personnalisée des risques de maladies chroniques basée sur l'ontologie. La modélisation personnalisée est un processus de création de modèle pour une seule personne, à partir de ses données personnelles et des informations disponibles dans l'ontologie. Un système d'inférence traductive neuro-floue avec normalisation pondérée des données est utilisé pour évaluer le risque personnalisé de maladie chronique. Cette approche vise à fournir un support pour de nouvelles découvertes grâce à l'intégration de la représentation ontologique pour construire un système expert afin d'identifier les gènes d'intérêt et les composants alimentaires pertinents.

Les systèmes d'aide à la décision clinique (CDSS) assistent les médecins dans leur travail quotidien, améliorant ainsi la qualité des soins prodigués à un patient. Ils les accompagnent dans le processus de prise de décision et leur propose des traitements adaptés [184]. L'utilisation de l'ontologie pour construire des systèmes d'aide à la décision basés sur la connaissance est largement adoptée. L'ontologie est la mieux adaptée pour encapsuler les concepts et les relations de termes associés au domaine médical. Il est adapté pour capturer les connaissances médicales de manière formelle, permettant de les partager et de les réutiliser chaque fois que nécessaire. Tous les concepts et relations détaillés dans les directives cliniques peuvent être mis en œuvre à l'aide du langage d'ontologie Web (OWL). Le mécanisme de raisonnement est vital dans tout système basé sur la connaissance. L'ontologie peut être raisonnée pour recommander le traitement approprié pour un patient en tenant compte de l'état de santé actuel du patient. Dans ce travail, un système d'aide à la décision basé sur une ontologie appelé OntoDiabetic a été développé pour évaluer les facteurs de risque et fournir des suggestions de traitement appropriées pour les patients diabétiques. Cet article se concentre sur la modélisation et la mise en œuvre de directives cliniques à l'aide des règles OWL2 et du processus de raisonnement dans le système OntoDiabetic. L'étude de cas est menée pour des patients présentant un risque de maladie cardiovasculaire manifeste, de néphropathie diabétique et d'hypertension dans les centres de santé primaires d'Oman.

[185] Cet article décrit une approche qui utilise une ontologie floue pour tenter à la fois d'améliorer la prédiction des maladies cardiovasculaires et de fournir une capacité prédictive personnalisée. L'objectif des auteurs est de suggérer que l'utilisation d'ontologies et en particulier d'ontologies floues peut être utile pour l'expression des connaissances sur les facteurs de risque et les résultats et que les approches d'ensembles flous de type 2 peuvent permettre la représentation de l'incertitude à la fois dans la collecte de données et la certitude diagnostique. En pratique, la combinaison de l'ontologie et des raisonneurs pourrait être utilisée pour prédire le résultat.

Identifier et distinguer les gènes moteurs du cancer parmi des milliers de mutations candidates reste un défi majeur. L'identification précise des gènes conducteurs et des mutations conductrices est essentielle pour faire avancer la recherche sur le cancer et personnaliser le traitement en fonction d'une stratification précise des patients. En raison de l'hétérogénéité génétique inter-tumorale, de nombreuses mutations conductrices au sein d'un gène se produisent à de faibles fréquences, ce qui rend difficile leur distinction des mutations non conductrices. [186] Les auteurs ont développé une nouvelle méthode pour identifier les gènes moteurs du cancer. Leur approche utilise plusieurs types d'informations complémentaires, en particulier les phénotypes cellulaires, les emplacements cellulaires, les fonctions et les phénotypes physiologiques du corps entier comme caractéristiques. Leur méthode repose sur l'apprentissage en profondeur sur des ontologies intégrées et des bases de connaissances biologiques, et souligne l'importance d'utiliser des ressources structurées et formalisées comme connaissances de base dans les modèles d'apprentissage automatique. Ils ont démontré que la méthode peut identifier avec précision les gènes connus responsables du cancer et distinguer leur rôle dans différents types de cancer. En plus de confirmer les gènes conducteurs connus, nous identifions plusieurs nouveaux gènes conducteurs candidats. Ils ont démontré l'utilité de notre méthode en validant ses prédictions dans le cancer du nasopharynx et le cancer colorectal à l'aide du séquençage de l'exome entier et du génome entier.

La raison d'une maladie peut varier d'une personne à l'autre. Cependant, l'ignorance est une cause fréquente de la plupart des maladies et est causée par un manque de conscience des symptômes que le corps humain indique. Ainsi, la prédiction de la maladie à un stade précoce devient une tâche importante. A l'aide de l'ontologie, les auteurs [187] ont fourni essentiellement des informations sur divers symptômes et maladies. L'ontologie développée comprend les maladies et leurs relations avec les symptômes, ainsi que les requêtes SPARQL (SPARQL et RDF Query Language) pour la prédiction des maladies. L'ontologie développée dans ce projet comporte deux étapes La première étape consiste à définir les classes, les sous-classes et les données, ainsi que les propriétés des objets. Ensuite, le fichier d'ontologie créé sera téléchargé sur le serveur apache jena fuseki. La deuxième partie consiste à exécuter des requêtes SPARQL à partir du serveur qui extrait les maladies avec des symptômes en fonction de la requête écrite.

Récemment, de grandes quantités de données ont été produites en raison des progrès réalisés dans les domaines de la biotechnologie et des sciences de la santé. Il comprend des informations cliniques et des données génétiques contenues dans les dossiers de santé électroniques (DSE). Par conséquent, il y avait un besoin de méthodes innovantes et efficaces pour représenter cette quantité de données. D'un autre côté, il est très important de détecter les syndromes, qui peuvent avoir une mauvaise influence sur la santé humaine en plus de mettre des charges financières sur leurs épaules, à un stade précoce pour éviter de nombreuses complications. Récemment, différentes techniques d'exploration de données en plus des techniques basées sur l'ontologie ont joué un grand rôle dans la construction de systèmes automatisés capables de détecter les syndromes de manière efficace et précise. Dans cet article [188], les auteurs ont couvert certains des efforts de recherche qui ont utilisé soit les techniques d'exploration de données, soit les techniques basées sur l'ontologie, soit les deux pour détecter les syndromes. De plus, un ensemble de techniques d'exploration de données bien connues, y compris les arbres de décision (J48), Naïve Bayes, le perceptron multicouche (MLP) et la forêt aléatoire (RF) a été évalué dans l'exécution de la tâche de

classification à l'aide d'un ensemble de données sur les maladies cardiaques accessible au public.

La détection de la maladie se fait en utilisant un algorithme de prédiction. Ici, un algorithme d'apprentissage automatique a été utilisé pour trouver la précision [189]. L'ensemble de données a été collecté auprès de certains hôpitaux et prétraité où les valeurs manquantes ont été reconstruites avant le processus de prédiction. En raison de l'énorme quantité d'informations dans le domaine de la santé, le résultat exact est le besoin de reconnaissance de la maladie et de services. Généralement, les données brutes sont de mauvaise qualité car elles ont l'exactitude, l'exhaustivité des champs d'enregistrement. De plus, il y aurait des expositions différentes selon les régions, les apparitions de certaines maladies, ce qui peut également affaiblir la prédiction de l'éclosion de la maladie. En utilisant le dossier de santé, ce système a bien réussi un taux d'exactitude est de 97%. Dans ce système proposé, les auteurs ont fourni une prédiction de diverses maladies qui se produisent grâce à l'utilisation de l'apprentissage automatique qui sera efficace.

Dans cette étude [190], les auteurs ont présenté une nouvelle approche pour la détection de faux comptes basée sur l'ontologie de domaine et les règles SWRL. L'ontologie contient les concepts pertinents liés à la découverte et aux données des comptes de profil Twitter. Les règles SWRL sont créées à partir de relations solides entre les concepts d'ontologie pour estimer le faux compte. Les fonctions de l'approche sont divisées en quatre tâches, à savoir la préparation des données, la classification et la relation, la construction d'ontologies et enfin l'application des règles SWRL pour déduire si le compte est faux ou non. Toutes les règles de détection et les connaissances pertinentes sont extraites de l'ontologie de domaine (FakAccOnt). Le raisonneur utilise les données et les règles du compte de profil pour tirer l'inférence et fournit la décision finale pour montrer que la nature du compte est fausse ou de confiance. Des expériences sont menées pour tester les performances de l'approche dans la reconnaissance de 3991 comptes Twitter. L'évaluation du système dépend des règles SWRL et des mesures d'évaluation standard, et les résultats montrent que l'approche proposée peut identifier correctement 2 768 des 2 805 faux comptes, soit une précision de 97,5 %. La principale contribution de cette étude est l'utilisation de l'ontologie et des règles SWRL pour détecter les comptes de profil Twitter en fonction de la fonction de compteur.

Cet article [191], vise à créer un modèle prédictif, qui aidera à l'attribution des commandes nouvellement reçues dans un réseau de fabrication. Le réseau de fabrication, qui est pris comme étude de cas dans cette recherche, se compose de plus de 300 petites entreprises de fabrication avec une société centrale en tant qu'intégrateur de gestion de projet. La méthodologie présente la cartographie d'un modèle d'ontologie basé sur PROSA (Product-Resource-Order-Staff Architecture) sur un arbre de décision, qui a été créé avec l'application Waikato Environment for Knowledge Analysis (WEKA). En outre, la méthodologie démontre également la formulation des règles du langage de règles du Web sémantique (SWRL) à partir de l'arbre de décision WEKA à l'aide de la programmation MATLAB. Le modèle a donné une précision de prédiction de 60,4 % en utilisant 8 fournisseurs, ce qui est considéré comme une base solide pour un modèle visant à illustrer l'idée d'attribution d'une commande nouvellement reçue.

Dans cet article [192], une nouvelle approche a été proposée pour détecter et classer les faux comptes sur les réseaux sociaux Twitter, en utilisant l'ingénierie ontologique. Les auteurs ont

modélisé une approche ontologique de la représentation des connaissances à travers le langage OWL, les règles SWRL et le raisonneur. Ils se sont concentrés sur les caractéristiques des profils qui pourraient être davantage traduites en axiomes. Le raisonneur a été utilisé pour exécuter toutes les requêtes d'ontologie OWL afin d'obtenir les réponses correctes aux requêtes. Le raisonneur Pellet a été utilisé comme classificateur pour la détection et la classification des comptes Twitter. L'approche proposée a été réalisée sur la base des métriques standard, en utilisant 3991 comptes de profil Twitter. Le système a correctement identifié 2758 des 2797 comptes comme étant de faux comptes avec un taux de précision de 97,5 %. Lors de la phase de classification, les résultats ont indiqué que 1672 faux comptes sur 1776 étaient classés comme spam bots, tandis que les 1016 comptes restants étaient classés comme faux abonnés avec un taux de précision de 96,1%.

Dans cet article [193], les auteurs ont introduit une nouvelle approche basée sur l'ontologie pour faciliter la maintenance prédictive dans l'industrie. L'approche proposée est une utilisation combinée des technologies de clustering flou et sémantique où les techniques de clustering flou sont utilisées pour apprendre la criticité des défaillances sur la base des données historiques de la machine, et les technologies sémantiques utilisent les résultats du clustering flou pour prédire le moment des défaillances et la criticité d'eux. En conséquence, une ontologie de domaine pour la modélisation des connaissances en maintenance prédictive est développée, et un ensemble de règles prédictives SWRL est proposé pour raisonner sur le temps et la criticité des pannes de machines. Une étude de cas sur un ensemble de données industrielles du monde réel est suivie pour évaluer l'utilité et l'efficacité de l'approche proposée.

Dans le papier [194], les auteurs ont présenté un système capable de prédire automatiquement si les patients développent une maladie coronarienne en fonction de leurs antécédents médicaux narratifs, c'est-à-dire du texte libre clinique. Bien que le texte libre des dossiers médicaux ait été utilisé dans plusieurs études pour identifier les facteurs de risque de maladie coronarienne, à leur connaissance, leur travail marque la première tentative de prédiction automatique du développement de la coronaropathie. Ils se sont attaqués à cette tâche sur un petit corpus de patients diabétiques. La taille de ce corpus rend important de limiter le nombre de caractéristiques afin d'éviter le sur justement. Ils ont proposé une approche guidée par l'ontologie de l'extraction de caractéristiques et l'ont comparée à deux techniques classiques de sélection de caractéristiques. Leur système atteint des performances de pointe de 77,4% de score F1.

La stéatose hépatique non alcoolique est une complication clinique courante. L'article [195] vise à développer un système de détection de la stéatose hépatique basé sur des connaissances basées sur une ontologie et des règles de détection extraites d'un algorithme d'arbre de décision. L'ontologie est créée pour représenter les connaissances liées aux patients et à la stéatose hépatique. En utilisant 43 règles SWRL et le moteur d'inférence Drool en ontologie, nous avons détecté des patients atteints de stéatose hépatique. La taille de l'ensemble de données de formation correspond à 70 % des données propres, y compris 580 dossiers médicaux électroniques de patients souffrant de maladies du foie. Après déduction des règles, le nombre de patients souffrant de stéatose hépatique en ontologie est le même que le modèle d'arbre de décision. L'article a validé le résultat généré par le modèle d'ontologie à travers les résultats du modèle d'arbre de décision.

II.4. Conclusion

De nos jours, l'apprentissage automatique est également utilisé par l'industrie de la santé pour faire progresser leurs techniques afin qu'ils puissent fournir de meilleurs services à leurs patients. Le système de prédiction des maladies prédit les maladies en fonction des symptômes du patient et également de certains médicaments couramment prescrits pour une maladie particulière. D'autre part, les ontologies peuvent soutenir le diagnostic des maladies, en particulier en raison de leur capacité inhérente à traiter l'interopérabilité sémantique. Par conséquent, de nouvelles approches basées sur l'ontologie doivent être intégrées, par exemple, la prédiction des maladies à l'aide de techniques d'ontologie et d'apprentissage automatique. De nombreux travaux de recherche ont été menés pour prédire les maladies en fonction des symptômes présentés par un individu à l'aide de l'ontologie et les algorithmes d'apprentissage automatique.

Dans ce chapitre, nous avons vu un aperçu de la littérature existante concernant les approches de prédiction des maladies basées sur le web sémantique et l'apprentissage automatique. Nous avons cité tous les travaux récents concernant la prédiction des maladies cardiovasculaire, le cancer du sein, et ainsi la prédiction des cas de COVID-19. Nous remarquons par ailleurs, que la majorité des recherches ont été basées sur l'apprentissage automatique, tandis que l'utilisation de l'ontologie est quasiment limitée. Dans la suite de ce manuscrit, nous présentons nos contributions qui s'intéressent à l'intégration de l'ontologie avec l'apprentissage automatique pour la prédiction des maladies.

Comme nous l'avons mentionné précédemment, cette revue de la littérature a volontairement été gardée générique. Dans les chapitres restants, chaque chapitre aura une revue de littérature plus spécifique.

CHAPITRE III - L'IMPACT DE L'ONTOLOGIE SUR LA PREDICTION DES MALADIES CARDIOVASCULAIRES PAR RAPPORT AUX ALGORITHMES D'APPRENTISSAGE AUTOMATIQUE

III.1. Introduction

Les maladies cardiovasculaires ont été considérées comme la maladie la plus grave et la plus mortelle chez l'humain. Le taux accru de maladies cardiovasculaires avec un taux de mortalité élevé entraîne un risque et un fardeau importants pour les systèmes de santé du monde entier. Les maladies cardiovasculaires sont plus fréquentes chez les hommes que chez les femmes, en particulier à un âge moyen ou avancé [196], bien qu'il existe également des enfants ayant des problèmes de santé similaires [197], [198].

Selon les données fournies par l'Organisation mondiale de la santé (OMS), un tiers des décès dans le monde sont causés par des maladies cardiaques. Les maladies cardiovasculaires causent la mort d'environ 17,9 millions[9] de personnes chaque année dans le monde [199]. La Société européenne de cardiologie (ESC) a rapporté que 26 millions d'adultes dans le monde ont reçu un diagnostic de maladie cardiaque et 3,6 millions sont identifiés chaque année. Environ la moitié de tous les patients diagnostiqués avec une maladie cardiaque meurent en seulement 1 à 2 ans et environ 3% du budget total des soins de santé est consacré au traitement des maladies cardiaques [127].

Pour prédire une maladie cardiaque, plusieurs tests sont nécessaires, ajouté à cela le manque d'expertise du personnel médical peut entraîner de fausses prédictions [200], d'autre part le diagnostic précoce peut être difficile. Le traitement chirurgical des maladies cardiaques est difficile, en particulier dans les pays en développement qui manquent de personnel médical qualifié ainsi que d'équipements de test et d'autres ressources nécessaires pour un diagnostic et des soins appropriés des patients souffrant de problèmes cardiaques [201]. Une évaluation précise du risque d'insuffisance cardiaque permettrait de prévenir les infarctus graves et d'améliorer la sécurité des patients [202].

Les algorithmes d'apprentissage automatique peuvent être efficaces pour identifier les maladies, lorsqu'ils sont formés sur des données appropriées. Les ensembles de données (datasets) sur les maladies cardiaques sont accessibles au public pour la comparaison des modèles de prédiction. L'introduction de l'apprentissage automatique et de l'intelligence artificielle aide les chercheurs à concevoir le meilleur modèle de prédiction en utilisant les grandes bases de données disponibles. Des études récentes portant sur les problèmes cardiaques chez les adultes et les enfants ont souligné la nécessité de réduire la mortalité liée aux maladies cardiovasculaires.

Étant donné que les « datasets » cliniques disponibles sont incohérents et redondants, un prétraitement approprié est une étape cruciale [203]. Il est essentiel de sélectionner les caractéristiques significatives qui peuvent être utilisées comme facteurs de risque dans les modèles de prédiction. Il faut veiller à sélectionner la bonne combinaison de caractéristiques

[9] https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

et les algorithmes d'apprentissage automatique appropriés pour développer des modèles de prédiction précis [204]. Il est important d'évaluer l'effet des facteurs de risque qui répondent aux trois critères comme la prévalence élevée dans la plupart des populations ; un impact significatif sur les maladies cardiaques indépendamment ; et ils peuvent être contrôlés ou traités pour réduire les risques.

Différents chercheurs ont inclus différents facteurs de risque ou caractéristiques lors de la modélisation des prédicteurs des maladies cardiovasculaires. Les caractéristiques utilisées dans le développement de modèles de prédiction des maladies cardiovasculaires dans différents travaux de recherche comprennent l'âge, le sexe, les douleurs thoraciques (cp), la glycémie à jeun (FBS) - un FBS élevé est lié au diabète [205], les résultats électrocardiographiques au repos (Restecg), l'exercice- angine induite (exang), dépression ST induite par l'exercice par rapport au repos (oldpeak), pente, nombre de vaisseaux principaux colorés par fluoroscopie (ca), état cardiaque (thal), fréquence cardiaque maximale atteinte (thalach), mauvaise alimentation, antécédents familiaux , le cholestérol (chol), l'hypertension artérielle, l'obésité, l'inactivité physique et la consommation d'alcool [206]–[210]. Des études récentes révèlent un besoin d'un minimum de 14 attributs pour rendre la prédiction précise et fiable [211].

Les chercheurs actuels ont du mal à combiner ces caractéristiques avec les techniques d'apprentissage automatique appropriées pour faire une prédiction précise des maladies cardiaques [212]. Les algorithmes d'apprentissage automatique sont plus efficaces lorsqu'ils sont entraînés sur des « datasets » appropriés [213]–[215]. Étant donné que les algorithmes reposent sur la cohérence des données d'entraînement et de test, l'utilisation de techniques de sélection de caractéristiques telles que l'exploration de données, la sélection de relief et LASSO (Least Absolute Shrinkage and Selection Operator) peut aider à préparer les données afin de fournir une prédiction plus précise. Une fois les caractéristiques pertinentes sélectionnées, des classificateurs et des modèles hybrides peuvent être appliqués pour prédire les chances d'apparition de la maladie. Les chercheurs ont appliqué différentes techniques pour développer des classificateurs et des modèles hybrides. Il existe encore un certain nombre de problèmes qui peuvent empêcher une prédiction précise des maladies cardiaques, comme des « datasets » médicales limités, la sélection de fonctionnalités, les applications d'algorithmes d'apprentissage automatique et un manque d'analyse approfondie.

Le Web sémantique est défini comme « une extension du Web actuel dans laquelle l'information est dotée d'une signification bien définie, ce qui permet aux ordinateurs et aux personnes de mieux travailler en coopération ». Il peut être utilisé pour organiser les connaissances en fonction de leur signification et permettre aux outils automatisés de vérifier les incohérences et d'extraire de nouvelles connaissances. [216] Aujourd'hui, de nombreux systèmes de soins de santé en ligne commencent à utiliser les technologies du web sémantique car beaucoup d'entre eux manquent de technologie de représentation des connaissances.

L'ontologie est une partie importante de l'architecture du Web sémantique. Il fournit une compréhension commune d'un domaine. Récemment, l'ontologie a été utilisée dans les soins de santé de différentes manières, telles que la représentation des connaissances du domaine, la fourniture de métadonnées pour les concepts et entités clés, permettant une description et

une récupération plus riches, facilitant l'échange et le partage, la personnalisation et la recommandation, etc... [217].

Dans ce chapitre, nous avons introduit une nouvelle approche consistant à fusionner l'apprentissage automatique et le Web Sémantique. D'une part, les algorithmes d'apprentissage automatique apprennent de manière autonome à effectuer une tâche ou à faire des prédictions à partir de données et améliorent leurs performances dans le temps, tandis que le Web sémantique fournit plusieurs formats d'affichage des données et des connaissances ontologiques de base. La fusion des deux nous a permis de construire un modèle basé sur une ontologie capable de prédire les maladies cardiovasculaires avec une grande précision. Nous avons établi un modèle ontologique de représentation des connaissances à travers le langage OWL, les règles SWRL et le raisonneur. Pour ce faire, nous générons les règles à partir de l'algorithme d'arbre de décision, puis nous les implémentons dans l'ontologie en utilisant le langage de règles pour le web sémantique (SWRL). Nous présentons également une analyse comparative entre les sept techniques de classification populaires et la classification d'apprentissage automatique basée sur l'ontologie, basée sur des paramètres soigneusement choisis tels que la Précision, l'Exactitude, le Rappel et la F-mesure, qui sont dérivés de la matrice de confusion.

III.2. Revue de littérature

Dernièrement, les chercheurs ont publié une quantité importante de recherches utilisant des approches d'apprentissage automatique pour détecter les personnes à risque de maladie cardiovasculaire en fonction des symptômes. Ces techniques se sont avérées impartiales et bénéfiques. L'application de l'intelligence artificielle et précisément les algorithmes d'apprentissage automatique ont gagné en popularité ces dernières années en raison de l'amélioration de la précision et de l'efficacité des prédictions [218], [219]. L'importance de la recherche dans ce domaine réside dans la possibilité de développer et de sélectionner des modèles avec la plus grande précision et efficacité [220]. Les modèles hybrides qui intègrent différents modèles d'apprentissage automatique avec des systèmes d'information (facteurs majeurs) sont une approche prometteuse pour la prédiction des maladies. Divers ensembles de données publiques disponibles sont appliqués.

Différentes approches sont appliquées dans le domaine de la prédiction des maladies cardiaques [221], telles que la machine à vecteurs de support (SVM), l'algorithme de forêt aléatoire (RF), le réseau de neurones artificiels (ANN), l'optimisation par essaims particulaires (PSO), l'algorithme génétique (GA), la classification naïve bayésienne (NB), la méthode des k plus proches voisins (KNN), et l'arbre de décision (DT) [222]–[224]. Dans cette partie, nous aborderons les plus récents d'entre eux.

Cette recherche [225] vise à évaluer les performances de divers algorithmes d'apprentissage automatique et à prévoir quel algorithme fonctionnerait mieux dans ce scénario. Les auteurs ont collecté des ensembles de données brutes, les ont prétraités et les ont testés pour faire une prédiction. Ils ont utilisé cinq types distincts d'algorithmes de classification pour prédire les maladies cardiaques dans cette procédure. Après avoir préparé les données, ils les ont passées à travers divers algorithmes de catégorisation pour voir leurs performances. Ils ont utilisé les algorithmes : méthode des k plus proches voisins, classification naïve bayésienne, arbre de

décision, algorithme de forêt aléatoire et la machine à vecteurs de support. Puis évaluer les performances de chaque algorithme en utilisant les valeurs d'exactitude, de précision, de rappel et de F-mesure. Certains algorithmes ont donné les meilleurs résultats ou certains ont donné les moins bons résultats dans certains cas. La classification naïve bayésienne a obtenu la meilleure précision pour l'ensemble de données utilisées. L'algorithme machine à vecteurs de support a également bien fonctionné, mais par rapport à l'algorithme de la classification naïve bayésienne, son résultat était médiocre. Ici, l'arbre de décision a mal fonctionné dans certains cas. L'algorithme de forêt aléatoire s'est également bien comporté car il a utilisé de nombreux arbres de décision pour surmonter le problème de surajustement. Le résultat de leur étude indique que la classification naïve bayésienne a exécuté le meilleur de tous les algorithmes, avec une précision de 83,96 %. Machine à vecteurs de support a fonctionné admirablement, avec une précision de 84,08 %, pratiquement identique à la classification naïve bayésienne. Bien qu'il ait une précision supérieure à la classification naïve bayésienne, ses performances sont médiocres en termes de précision, de rappel et de F-mesure. Finalement, la classification naïve bayésienne a obtenu de bons résultats, ce qui peut être utilisé pour prédire les maladies cardiaques.

Cette étude [226] visait à identifier les classificateurs d'apprentissage automatique avec la plus grande précision à des fins de diagnostic. Plusieurs algorithmes d'apprentissage automatique supervisés ont été appliqués et comparés pour leurs performances et leur précision dans la prédiction des maladies cardiaques. Les auteurs ont collecté un ensemble de données sur les maladies cardiaques, prétraité si nécessaire, puis ont été exécutés pour mieux comprendre l'ensemble de données. Ensuite, ils ont appliqué six algorithmes d'apprentissage automatique, AdaboostM1, Régression Logistique, Perceptron Multicouche, méthode des k plus proches voisins, arbre de décision et l'algorithme de forêt aléatoire, et ont évalué leurs prédictions en fonction de l'exactitude, de la sensibilité, de la spécificité, des statistiques kappa, de la précision, du rappel, de la F-mesure et du MCC, de la courbe ROC et Courbe Précision-Rappel. Les scores d'importance des caractéristiques pour chaque caractéristique ont été estimés pour tous les algorithmes appliqués, à l'exception du Perceptron Multicouche et la méthode des k plus proches voisins. Toutes les caractéristiques ont été classées en fonction du score d'importance pour trouver celles qui donnent des prédictions élevées sur les maladies cardiaques. Cette étude a révélé les bonnes performances de tous les algorithmes appliqués, où la méthode des k plus proches voisins, l'arbre de décision et l'algorithme de forêt aléatoire ont montré les meilleures performances avec une précision de 100 %, ce qui indique qu'ils sont les plus efficaces pour prédire les maladies cardiaques. Nous avons également estimé l'importance des caractéristiques et les valeurs de coefficient de tous les algorithmes appliqués, à l'exception du perceptron multicouche et la méthode des k plus proches voisins, car ces deux algorithmes n'ont généré aucun score d'importance des caractéristiques ni aucune valeur de coefficient. Cette analyse a identifié des caractéristiques hautement prédictives pour la détection des maladies cardiaques qui présentent une utilité potentielle pour les cliniciens cherchant à prédire l'apparition de maladies cardiaques chez leurs patients.

Cet article d'enquête [227] est une consolidation des travaux réalisés dans le domaine de la prédiction des maladies cardiovasculaires à l'aide de techniques d'apprentissage automatique et en profondeur. Il compare et rapporte les différents modèles de classification, d'exploration de données, d'apprentissage automatique et d'apprentissage en profondeur utilisés pour la

prédiction des maladies cardiovasculaires. L'enquête est organisée en trois volets : techniques de classification et d'exploration de données pour les maladies cardiovasculaires, modèles d'apprentissage automatique pour les maladies cardiovasculaires et modèles d'apprentissage en profondeur pour la prédiction des maladies cardiovasculaires. Les mesures de performance utilisées pour rendre compte de l'exactitude, l'ensemble de données utilisé pour la prédiction et la classification, et les outils utilisés pour chaque catégorie de ces techniques sont également compilés et rapportés dans cette enquête.

Un diagnostic précoce est essentiel pour prédire les maladies qui affectent le cœur humain et conduisent les patients à vivre une autre période de la vie [228]. Dans ce contexte, les auteurs présentent deux méthodes de diagnostic précoce prédisant si les individus ont ou non une maladie cardiaque, la machine à vecteurs de support et le réseau de neurones artificiels (ANN). Les données médicales sont extraites de la base de données Machine Learning « Repository » de l'Université de Californie à Irvine (UCI) et contiennent les rapports de 170 personnes. Les données des 170 individus sont classées en 90 ensembles de données d'apprentissage et 80 ensembles de données de test. La précision et la sensibilité sont mesurées pour les deux techniques dans la prédiction de chaque maladie. Les résultats de l'enquête confirment que l'exécution optimale est la technique de la machine à vecteurs de support. Il donne des résultats de prédiction de haute précision. Quant aux performances de la technique des réseaux de neurones artificiels à propagation directe, elles sont acceptables.

Shah et al [229] présentent divers attributs liés aux maladies cardiaques et le modèle sur la base d'algorithmes d'apprentissage supervisé tels que la classification naïve bayésienne, arbre de décision, la méthode des k plus proches voisins et algorithme de forêt aléatoire. Il utilise l'ensemble de données existant de la base de données Cleveland du référentiel UCI des patients atteints de maladies cardiaques. L'ensemble de données comprend 303 instances et 76 attributs. Sur ces 76 attributs, seuls 14 attributs sont pris en compte pour les tests, importants pour justifier les performances des différents algorithmes. Ce document de recherche vise à envisager la probabilité de développer une maladie cardiaque chez les patients. Les résultats montrent que le score de précision le plus élevé est obtenu avec la méthode des k plus proches voisins.

Les auteurs l'étude [91] proposent un modèle qui intègre différentes méthodes pour obtenir une prédiction efficace des maladies cardiaques. Ils ont utilisé des méthodes efficaces de collecte de données, de prétraitement des données et de transformation des données pour créer des informations précises pour le modèle entraineur. Ils ont utilisé un ensemble de données combiné (Cleveland, Long Beach VA, Suisse, Hongrois et Stat log). Les caractéristiques appropriées sont sélectionnées à l'aide des techniques Relief et Least Absolute Shrinkage and Selection Operator (LASSO). De nouveaux classificateurs hybrides tels que Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), K-Nearest Neighbors Bagging Method (KNNBM), AdaBoost Boosting Method (ABBM) et Gradient Boosting Boosting Method (GBBM) sont développés en intégrant les classificateurs traditionnels avec des méthodes d'ensachage et de renforcement, qui sont utilisés dans le processus de l'entraînement. Ils ont également instrumenté certains algorithmes d'apprentissage automatique pour calculer l'exactitude (ACC), la sensibilité (SEN), le taux d'erreur, la précision (PRE) et le score F1 (F1) de leur modèle, ainsi que la valeur prédictive

négative (NPR), les faux positifs taux (FPR) et taux de faux négatifs (FNR). Sur la base de l'analyse des résultats, ils concluent que le modèle proposé a produit la plus grande précision en utilisant les méthodes de sélection des caractéristiques RFBM et Relief (99,05%).

La recherche [230] étudie la performance des techniques d'apprentissage automatique pour prédire la probabilité de maladies cardiovasculaires. Les auteurs ont proposé un modèle pour diagnostiquer la probabilité qu'un individu ait une maladie cardiovasculaire en utilisant des modèles d'apprentissage automatique. Les expériences ont été exécutées à l'aide de sept algorithmes, à savoir la régression logistique, l'arbre de décision, la forêt aléatoire, la classification naïve bayésienne, la méthode des k plus proches voisins, la machine à vecteurs de support, le perceptron multicouche et un ensemble de données public sur les maladies cardiovasculaires a été utilisé pour entraîner les modèles. Un test « chi-square » a été utilisé pour identifier les caractéristiques les plus importantes pour prédire les maladies cardiovasculaires. Les résultats de l'expérience ont montré que le perceptron multicouche donne la plus grande précision de prédiction de la maladie à 87,23%.

Dans cet article [98], les auteurs ont proposé une nouvelle méthode appelée Hybrid Random Forest with Linear Model (HRFLM), qui vise à trouver des caractéristiques significatives en appliquant des techniques d'apprentissage automatique permettant d'améliorer la précision de la prédiction des maladies cardiovasculaires. Le modèle de prédiction est introduit avec différentes combinaisons de caractéristiques et plusieurs techniques de classification connues. Ils ont produit un niveau de performance amélioré avec un niveau de précision de 88,7% grâce au modèle de prédiction des maladies cardiaques en se basant sur la forêt aléatoire hybride et un modèle linéaire (HRFLM). Les résultats de l'expérience montrent que leur méthode hybride proposée à une plus grande capacité à prédire les maladies cardiaques par rapport aux méthodes existantes.

L'industrie de la santé a constaté que l'apprentissage automatique est une technique de prise de décision précieuse et précise dans la collecte de données produites en grande quantité. Les systèmes d'aide à la décision médicale développés se sont avérés efficaces sur la base des logiciels et des différents algorithmes proposés par de nombreux chercheurs. [231] Cette étude se fait sur la base des différentes techniques utilisant les différents algorithmes et leur analyse des performances. Le modèle de prédiction a été introduit avec plusieurs caractéristiques combinées, et parmi les multiples méthodes et figuraient d'autres techniques de classification. De nombreuses méthodes existantes ont été discutées, parmi lesquelles le niveau de précision a été trouvé à 88,7 % en utilisant la technique de la forêt aléatoire hybride avec un modèle linéaire (HRFLM).

Pour la prédiction des problèmes cardiovasculaires une analyse qui été menée [232] utilisant les outils Weka pour la prédiction en se basant sur des algorithmes d'extraction de données comme l'optimisation minimale séquentielle (SMO), le perceptron multicouche (MLP) et la forêt aléatoire et le réseau de Bayes. Les données collectées combinent les résultats de précision de la prédiction, la courbe des caractéristiques de fonctionnement du récepteur (ROC) et la valeur PRC. Les performances des algorithmes de réseau de Bayes (94,5 %) et de forêt aléatoire (94 %) indiquent des performances optimales plutôt que les méthodes d'optimisation séquentielle minimale (SMO) et de perceptron multicouche (MLP).

Dans cet article [233], une étude a été présentée pour aider les spécialistes et les médecins irakiens à enquêter sur les problèmes cardiaques via le logiciel Weka en se concentrant sur

quatre techniques de classification d'exploration de données (1BK, J48, Naïve Bayes et REPTREE). Les tests de précision prédictive, la courbe ROC et la valeur AUC sont calculés sur la base d'un ensemble de données compilées reçues de l'hôpital de la ville médicale de Bagdad et de l'hôpital d'Ibn al-Bitar. La performance de la technique J48 (94,5%) indique une performance optimale basée sur SMO.

Dans le cadre du travail [234], l'état de l'art a été étudié pour identifier les principaux algorithmes prédictifs. En outre, ces algorithmes, à savoir la machine à vecteurs de support (SVM), la classification naïve bayésienne (NB), l'arbre de décision (DT), l'algorithme de forêt aléatoire (RF), le réseau de neurones artificiels (ANN), La régression logistique (LR), AdaBoost et la méthode des k plus proches voisins (k-NN) analysés par rapport aux deux ensembles de données sur le logiciel open source WEKA. Ce travail a utilisé deux ensembles de données structurés similaires, à savoir l'ensemble de données Statlog et l'ensemble de données Cleveland. Pour le prétraitement des ensembles de données, les valeurs manquantes ont été remplacées par la valeur moyenne et plus tard, une validation croisée de 10 fois a été utilisée pour l'évaluation. Le résultat de l'analyse des performances a montré que la machine à vecteurs de support surpasse les autres algorithmes par rapport aux deux ensembles de données. Machine à vecteurs de support a montré une précision de 84,156 % par rapport à l'ensemble de données Cleveland et de 84,074 % par rapport à l'ensemble de données Statlog. LR a montré une zone ROC de 0,9 par rapport aux deux ensembles de données. Les résultats des travaux aideront les établissements de santé à comprendre l'importance et l'utilisation des algorithmes prédictifs pour la prédiction automatique des maladies cardiovasculaires en fonction des symptômes.

Dans cet article [235], une application basée sur Python est développée pour la recherche en soins de santé car elle est plus fiable et permet de suivre et d'établir différents types d'applications de surveillance de la santé. Les auteurs ont décrit les principales phases du développement d'applications : collecte de bases de données, réalisation d'une régression logistique et évaluation des attributs de l'ensemble de données. Un algorithme de classificateur de forêt aléatoire est développé pour identifier les maladies cardiaques avec une plus grande précision. L'analyse des données est nécessaire pour cette application, qui est considérée comme importante en raison de son taux de précision d'environ 83 % sur les données d'apprentissage. Ensuite, ils ont discuté de l'algorithme de classification aléatoire des forêts, y compris les expériences et les résultats, qui fournissent de meilleures précisions pour les diagnostics de recherche.

Dans cette recherche [236], différentes techniques d'apprentissage automatique renommées : réseaux de neurones artificiels, machines à vecteurs de support, naïve bayes, arbres de décision et forêts aléatoires ont été étudiées pour aider à construire, comprendre et interpréter différents modèles de diagnostic des maladies cardiaques. Le modèle des réseaux de neurones artificiels a montré la meilleure précision de 84,25 % par rapport aux autres modèles. De plus, il a été constaté que malgré certains modèles conçus ayant des précisions plus élevées que d'autres, il peut être plus sûr de choisir un modèle de précision inférieure comme conception finale pour cette étude. Ce sacrifice était essentiel pour s'assurer qu'un modèle plus transparent et plus fiable est utilisé dans le processus de diagnostic des maladies cardiaques. Cette validation de la transparence a été effectuée à l'aide d'une nouvelle mesure

suggérée : l'indice de coût de classement des fonctionnalités. L'utilisation de cet indice a montré des résultats prometteurs en indiquant clairement quel modèle d'apprentissage automatique a un équilibre entre précision et transparence.

Cet article [237] propose un nouveau modèle hybride d'apprentissage en profondeur pour la prédiction des maladies cardiaques à l'aide d'un réseau neuronal récurrent (RNN) avec la combinaison de plusieurs unités récurrentes fermées (GRU), d'une mémoire longue à court terme (LSTM) et d'un optimiseur Adam. Ce modèle proposé a abouti à une précision exceptionnelle de 98,6876%, ce qui est le plus élevé du modèle existant de RNN. Le modèle a été développé en Python 3.7 en intégrant RNN dans plusieurs GRU qui fonctionnent dans Keras et Tensorflow en tant que backend pour le processus d'apprentissage en profondeur, pris en charge par diverses bibliothèques Python. Les modèles existants récents utilisant RNN ont atteint une précision de 98,23 % et le réseau neuronal profond (DNN) a atteint 98,5 %. Les inconvénients communs des modèles existants sont la faible précision due à la construction complexe du réseau neuronal, le nombre élevé de neurones avec redondance dans le modèle de réseau neuronal et les ensembles de données de déséquilibre de Cleveland. Des expériences ont été menées avec divers modèles personnalisés, où les résultats ont montré que le modèle proposé utilisant RNN et plusieurs GRU avec la technique de suréchantillonnage minoritaire synthétique (SMOTe) a atteint le meilleur niveau de performance. Il s'agit du résultat le plus précis pour RNN utilisant les ensembles de données de Cleveland et très prometteur pour faire une prédiction précoce des maladies cardiaques pour les patients.

Cet article [238] propose un système d'aide à la décision (DSS) pour diagnostiquer les maladies cardiovasculaires. Il utilise des approches d'apprentissage en profondeur qui classent les signaux d'électrocardiogramme (ECG). Ainsi, une architecture de réseau neuronal basée sur la mémoire à long terme (LSTM) à deux étages, ainsi qu'un prétraitement adéquat des signaux ECG, est conçue comme un système d'aide au diagnostic pour la détection de l'arythmie cardiaque basée sur une analyse du signal ECG. Ce système de diagnostic des maladies cardiovasculaires basé sur l'apprentissage profond (à savoir « DLCVD ») est conçu pour répondre à des exigences de performance plus élevées en termes de précision, de spécificité et de sensibilité. Celui-ci doit également être capable d'une classification en ligne en temps réel. Les résultats expérimentaux utilisant la base de données sur les arythmies du Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) montrent que le DLCVD a conduit à des performances exceptionnelles.

III.3. Méthodologie

Dans cette section, nous présentons les méthodologies et les matériaux utilisés, ainsi que le flux de travail expérimental, la description de l'ensemble de données, les algorithmes d'apprentissage automatique, le modèle d'ontologie et les métriques d'évaluation. Le flux de travail expérimental de cette analyse comparative est illustré dans la Figure III.1.

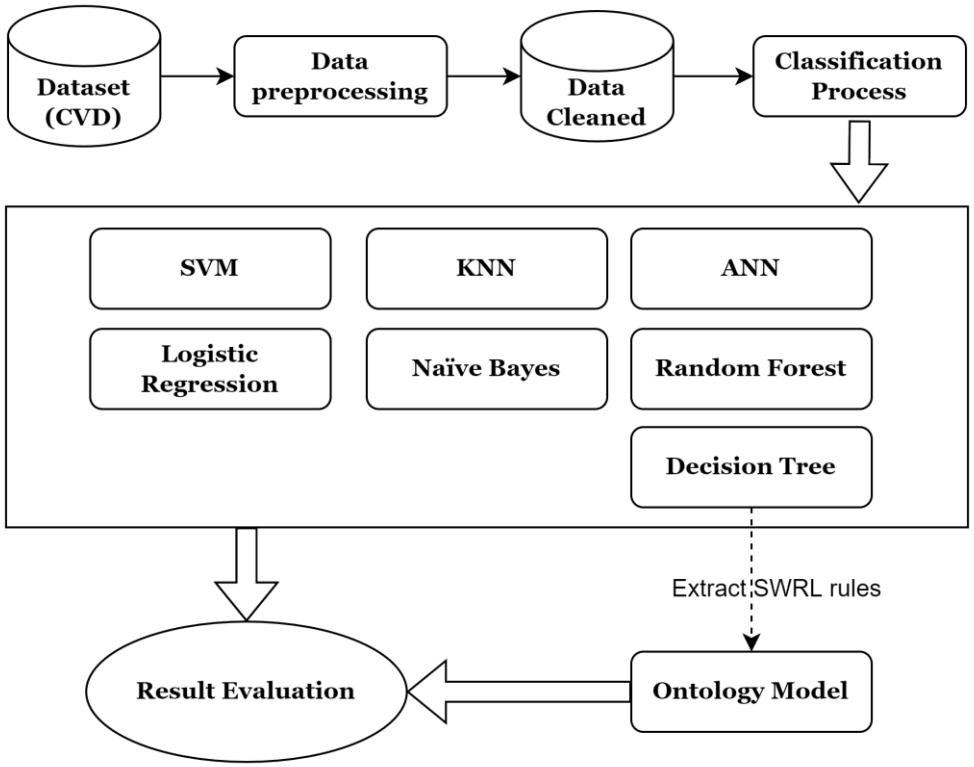

Figure III.1 - Flux de travail expérimental.

III.3.1. Prétraitement des données

L'ensemble de données utilisé est « Cardiovascular Disease dataset » depuis le site Web de Kaggle[10], il se compose de 70 000 instances et de 12 attributs (11 attributs et le dernier est l'attribut cible). Table III.1 donne une explication détaillée de toutes les caractéristiques de l'ensemble de données.

Table III.1 - Informations sur les caractéristiques de l'ensemble de données.

Attribut	*Description*
1- age	L'âge du patient (jours)
2- height	La taille du patient (cm)
3- weight	Le poids du patient (Kg)
4- gender	Le sexe du patient (Homme ou Femme)
5- ap_hi	La pression artérielle systolique
6- ap_lo	Pression sanguine diastolique
7-cholesterol	Le cholestérol du patient (1 : normal, 2 : supérieur à la normale, 3 : bien supérieur à la normale)

[10] https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

8- gluc	Glucose (1 : normal, 2 : supérieur à la normale, 3 : bien supérieur à la normale)
9- smoke	Le patient fume ou non (binaire)
10- alco	Le patient prend ou non de l'alcool (binaire)
11- active	Le patient est actif ou non (binaire)
12- cardio	Variable cible (0 ou 1).

Pour créer un classificateur d'apprentissage automatique efficace, nous devons toujours commencer par nettoyer les données, normaliser les fonctionnalités, transformer les fonctionnalités et même créer de nouvelles fonctionnalités à partir de l'ensemble de données. L'ensemble de données utilisé contient 24 instances similaires, après suppression des instances dupliquées, il reste 69976 instances où 35004 représente l'absence de maladie cardiovasculaire et 34972 représente la présence de cette dernière.

Nous n'avons ajouté aucune nouvelle fonctionnalité comme l'indice de masse corporelle (IMC) pour la raison qu'il n'y a pas beaucoup de différence en termes de résultats obtenus. Nous tenons à vous informer qu'afin de fournir une comparaison équitable des résultats de classification obtenus, nous n'avons utilisé aucune méthode de sélection de fonctionnalités ou d'amélioration des performances.

III.3.2. Algorithmes d'apprentissage automatique

Nous avons utilisé le logiciel WEKA[11] pour tous les algorithmes d'apprentissage automatique afin de prédire la maladie. WEKA (Waikato Environment for Knowledge Analysis) est une application logicielle gratuite et open-source conçue pour résoudre une série de problèmes d'exploration de données. Le Framework permet la mise en œuvre de plusieurs algorithmes d'analyse de données et fournit une API pour appeler des algorithmes intégrés à partir d'une application particulière par le langage de programmation JAVA. Il fournit une variété d'outils pour la classification, la régression, le regroupement, la suppression des fonctionnalités non pertinentes, la création de règles associées et la visualisation de l'ensemble de données.

Nous avons utilisé les sept classificateurs les plus utilisés pour classer les ensembles de données (Arbre de décision, Forêt aléatoire, Régression logistique, Réseau de neurones artificiels, classification naïve bayésienne, machine à vecteurs de support, méthode des k plus proches voisins). De plus, nous avons utilisé deux modes d'options de test : la validation croisée 10 fois et la répartition en pourcentage (entraînement fractionné à 60 %, le reste pour le test) dans le but d'enrichir l'étude.

Parmi les sept classificateurs utilisés, nous avons choisi l'algorithme d'arbre de décision pour implémenter les règles générées dans le modèle d'ontologie. L'algorithme arbre de décision fait partie de la famille des algorithmes d'apprentissage supervisé utilisés dans les statistiques, l'exploration de données et l'apprentissage automatique [239]. L'approche de l'arbre de décision, contrairement à d'autres algorithmes d'apprentissage supervisé, peut également être

[11] https://www.cs.waikato.ac.nz/ml/weka/

utilisée pour résoudre des problèmes de régression et de classification. Nous avons sélectionné l'algorithme d'arbre de décision pour de nombreuses raisons, le résultat de l'arbre de classification est plus facile à comprendre et à interpréter, et il prend en charge plusieurs types de données tels que numériques, nominaux, catégoriels, etc. L'objectif de l'utilisation d'un arbre de décision est de construire un modèle d'entraînement qui peut prédire la classe ou la valeur de la variable cible en apprenant des règles de décision de base à partir de données passées (données d'entraînement).

Figure III.2 illustre le résultat de la classification de l'arbre de décision et la Figure III.3 fournit un extrait de la sortie de l'arbre de décision (nous avons obtenu 584 feuilles) qui sera utilisé pour générer des règles SWRL qui seront utilisées dans le modèle d'ontologie.

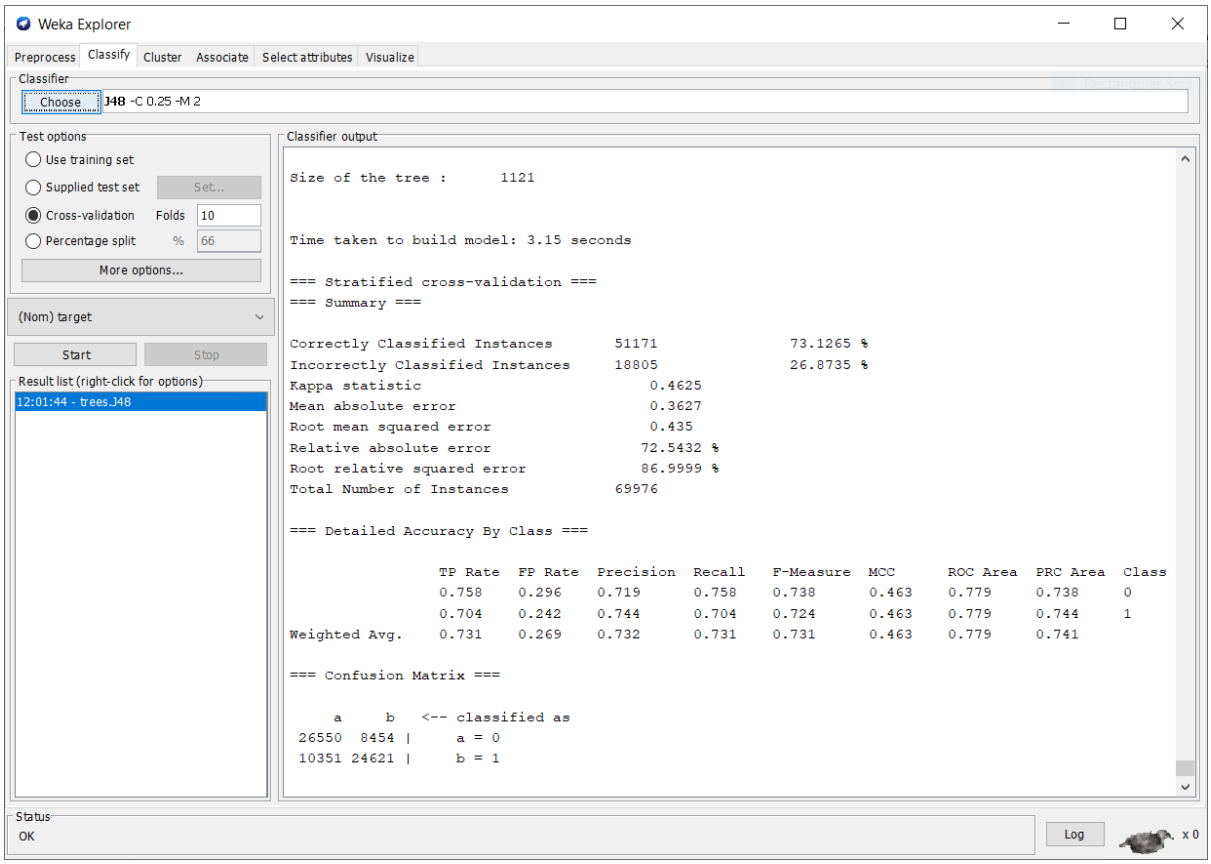

Figure III.2 - Résultat de la classification de l'arbre de décision.

```
J48 pruned tree
------------------

ap_hi <= 129
|   age <= 19931
|   |   cholesterol = 1
|   |   |   ap_hi <= 118
|   |   |   |   ap_hi <= 24
|   |   |   |   |   ap_hi <= 12
|   |   |   |   |   |   height <= 167: 0 (41.0/5.0)
|   |   |   |   |   |   height > 167
|   |   |   |   |   |   |   height <= 175
|   |   |   |   |   |   |   |   ap_hi <= 11
|   |   |   |   |   |   |   |   |   weight <= 68.5: 0 (3.0)
|   |   |   |   |   |   |   |   |   weight > 68.5: 1 (2.0)
|   |   |   |   |   |   |   |   ap_hi > 11: 1 (14.0/4.0)
|   |   |   |   |   |   |   height > 175: 0 (8.0/1.0)
|   |   |   |   |   ap_hi > 12: 1 (18.0/2.0)
|   |   |   |   ap_hi > 24
|   |   |   |   |   ap_lo <= 115: 0 (7484.0/1107.0)
|   |   |   |   |   ap_lo > 115
|   |   |   |   |   |   ap_hi <= 85
|   |   |   |   |   |   |   age <= 18428: 0 (14.0/1.0)
|   |   |   |   |   |   |   age > 18428: 1 (2.0)
|   |   |   |   |   |   ap_hi > 85
|   |   |   |   |   |   |   ap_lo <= 2088: 1 (23.0/4.0)
|   |   |   |   |   |   |   ap_lo > 2088: 0 (2.0)
|   |   |   ap_hi > 118
|   |   |   |   active <= 0
|   |   |   |   |   weight <= 82.5
|   |   |   |   |   |   gluc = 1: 0 (2220.0/597.0)

            ●●●

|   |   ap_lo <= 68
|   |   |   ap_hi <= 240
|   |   |   |   smoke <= 0: 1 (118.0/32.0)
|   |   |   |   smoke > 0
|   |   |   |   |   alco <= 0: 0 (8.0/1.0)
|   |   |   |   |   alco > 0: 1 (3.0)
|   |   |   ap_hi > 240
|   |   |   |   weight <= 79: 0 (10.0)
|   |   |   |   weight > 79: 1 (2.0)
|   |   ap_lo > 68: 1 (19270.0/3093.0)

Number of Leaves  :     584

Size of the tree :      1121
```

Figure III.3 - Extrait de la sortie de l'arbre de décision.

III.3.3. Modèle d'ontologie

Dans cette étape, l'ingénierie ontologique est utilisée pour construire l'ontologie et la représentation des connaissances, puis nous avons implémenté les règles de l'arbre de décision en les convertissant en raisonneur basé sur des règles du langage de règles du Web sémantique qui est utilisé pour détecter l'absence ou la présence de maladies cardiovasculaires.

III.3.3.1. Construction de l'ontologie

Protégé est utilisé pour construire l'ontologie, c'est une plateforme open-source qui offre une suite d'outils à une communauté d'utilisateurs croissante pour construire des modèles de domaine et des applications basées sur la connaissance avec des ontologies [240]. La Figure III.4 illustre la représentation graphique de l'ontologie que nous avons adopté pour notre cas. Nous avons créé deux classes principales « Patient » et « Diagnostic », deux sous-classes (absence et présence) de la classe mère Diagnostic, et quatre sous-classes (TP, TN, FN, FP) de la classe mère « PatientEvaluationMetrics » qui est une sous-classe de « Patient ». De plus, les propriétés des données sont les mêmes que les attributs de l'ensemble de données expliqués dans la Table III.1.

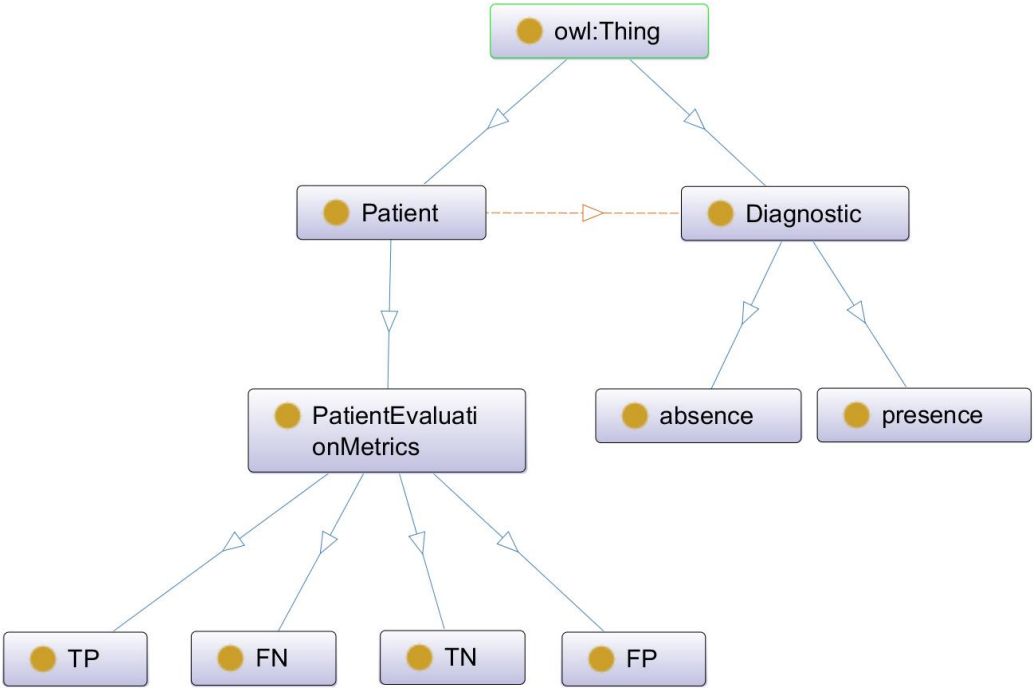

Figure III.4 - Représentation graphique de l'ontologie.

III.3.3.2. Propriétés des données et instances

Les propriétés de données utilisées dans l'ontologie sont les mêmes attributs présentés dans la Table III.1 qui sont utilisés pour construire des modèles d'algorithmes d'apprentissage automatique. La Figure III.5 illustre les propriétés des données.

A l'aide du Cellfie plugin du logiciel Protégé nous avons importé le même ensemble de données utilisé dans Weka. Ce plugin permet d'importer des données de feuille de calcul ou CSV dans des ontologies OWL.

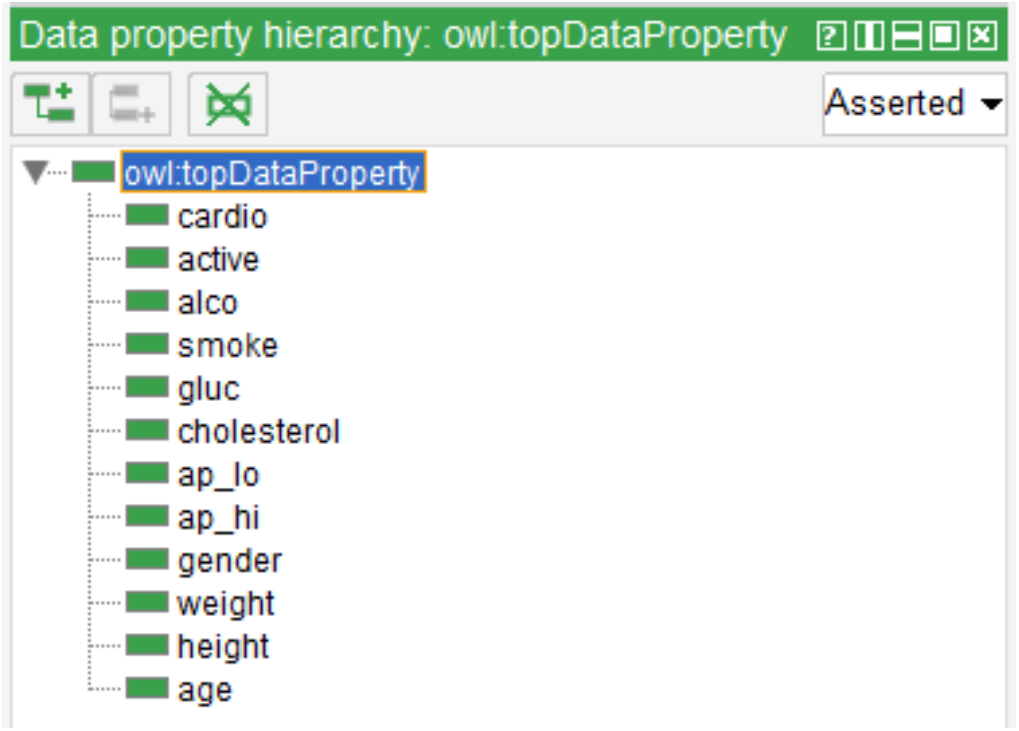

Figure III.5 - Propriétés des données.

III.3.3.3. Règles de langage du Web sémantique (SWRL)

Après avoir créé et rempli l'ontologie avec l'ensemble de données, dans cette étape, nous devons déterminer les règles du langage sémantique du Web pour le raisonnement. À cette fin, le langage de programmation Java est utilisé pour convertir les règles extraites de l'arbre de décision, comme illustré à la Figure III.3. Chaque feuille de l'arbre a été extraite en tant que règle SWRL unique à l'aide du programme Java. Par exemple, considérons la règle SWRL tirée de la première ligne de la Figure III.3.

Une feuille de l'algorithme de l'arbre de décision :

If cholesterol = 2 && alco ≤ 0 && smoke ≤ 0 && active ≤ 0 && weight ≤ 72 && ap_lo ≤ 85 && height ≤ 169 THEN put the patient in presence

SWRL obtenu :

Patient(?pt) ^ cholesterol(?pt, ?CH) ^ swrlb:equal(?CH, '2'^^xsd:decimal) ^ alco(?pt, ?AC) ^ swrlb:lessThanOrEqual(?AC, '0'^^xsd:decimal) ^ smoke(?pt, ?S) ^ swrlb:lessThanOrEqual(?S, '0'^^xsd:decimal) ^ active(?pt, ?A) ^ swrlb:lessThanOrEqual(?A, '0'^^xsd:decimal) ^ weight(?pt, ?W) ^ swrlb:lessThanOrEqual(?W, '72'^^xsd:decimal) ^ ap_lo(?pt, ?AL) ^ swrlb:lessThanOrEqual(?AL, '85'^^xsd:decimal) ^ height(?pt, ?H) ^ swrlb:lessThanOrEqual(?H, '169'^^xsd:decimal) → presence

III.3.3.4. Raisonneur Pellet

Nous avons utilisé le raisonneur Pellet, qui fournit des capacités plus directes pour travailler avec les règles OWL et SWRL, pour exécuter les règles SWRL et déduire de nouveaux axiomes d'ontologie. Pellet utilise l'ensemble de données et les règles SWRL pour induire

l'inférence et rend la décision finale quant à l'absence ou à la présence d'une maladie cardiovasculaire. Les résultats du classificateur d'ontologie sont rapportés dans la section suivante.

III.4. Évaluation

En apprentissage automatique, la mesure des performances est une tâche essentielle. Il est essentiel de choisir les bonnes métriques pour évaluer le modèle d'apprentissage automatique. Par conséquent, les métriques sont utilisées pour déterminer comment les performances des algorithmes d'apprentissage automatique sont mesurées et comparées.

Différentes mesures de performance sont utilisées pour évaluer les algorithmes d'apprentissage automatique tels que l'exactitude, la précision, le rappel, la F-mesure, la ROC area, la statistique Kappa, l'erreur quadratique moyenne racine, l'erreur quadratique relative racine, etc.

Presque toutes les mesures de performance sont dérivées de la matrice de confusion et des chiffres qu'elle contient. La matrice de confusion est l'une des mesures les plus intuitives et les plus simples pour déterminer l'exactitude et la précision du modèle. Il est utilisé pour les problèmes de classification avec deux ou plusieurs types de classes en sortie.

La matrice de confusion est un tableau à deux dimensions ("Actual" et "Predicted"), et des ensembles de "classes" dans les deux dimensions. Les classifications réelles sont des colonnes et les prévisions sont des lignes. Pour mieux comprendre ce qu'est la matrice de confusion et ce qu'elle représente, prenons un exemple concret de notre étude où nous prédisons si un patient est cardiaque ou non (1 : test positif ; 0 : test négatif). La Figure III.6 illustre les détails de la matrice de confusion et la Table III.2 décrit les termes associés à la matrice de confusion.

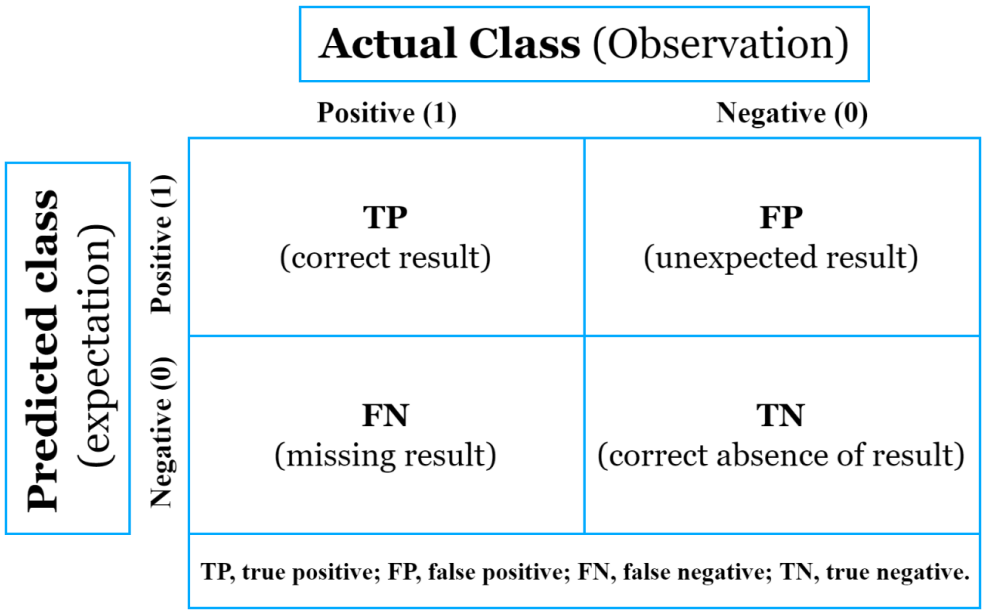

Figure III.6 - Détails de la matrice de confusion.

Table III.2 - Termes associés à la matrice de confusion.

Termes	*Description*
True Positives (TP)	Vrais positifs : représente le nombre d'individus malades avec un test positif.
True Negatives (TN)	Faux positifs : Représente le nombre d'individus non malades avec un test positif.
False Positives (FP)	Faux négatifs : Représente le nombre d'individus malades avec un test négatif.
False Negatives (FN)	Vrais négatifs : Représente le nombre d'individus non malades avec un test négatif.

Diverses mesures peuvent être dérivées d'une matrice de confusion telle que l'exactitude, la précision, le rappel et la F-mesure. La meilleure valeur d'exactitude, de précision et de rappel est de 1.0, tandis que la pire est de 0.0. La Figure III.7 illustre comment les calculer à partir de la matrice de confusion.

- **Exactitude (ACC) :**

Le facteur ACC est calculé comme le nombre de toutes les prédictions correctes divisé par le nombre total de l'ensemble de données, qui est le nombre de patients qui sont identifiés correctement au total dans notre cas.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Précision (PREC) :**

Ce facteur est calculé comme le nombre de prédictions positives correctes divisé par le nombre total de prédictions positives.

$$PREC = \frac{TP}{TP + FP}$$

- **Rappel (REC) :**

REC est calculé comme le nombre de prédictions positives correctes divisé par le nombre total de positifs. Il représente les patients pertinents qui ont été correctement détectés. Il est également appelé sensibilité ou taux de vrais positifs (TPR).

$$REC = \frac{TP}{TP + FN}$$

- **F-mesure**

Appelé aussi F-score, est une moyenne harmonique de précision et de rappel, il fournit la qualité de la prédiction.

$$\text{F-Measure} = 2 * \frac{PREC * REC}{PREC + REC}$$

- **ROC – AUC Area :**

La courbe AUC - ROC est une mesure de performance pour les problèmes de classification à différents réglages de seuil. ROC est une courbe de probabilité et AUC représente le degré ou la mesure de séparabilité. Il indique à quel point le modèle est capable de faire la distinction entre les classes. Plus l'AUC est élevée, plus le modèle prédit les classes 0 comme 0 et les classes 1 comme 1. Par analogie, plus l'AUC est élevée, plus le modèle est efficace pour distinguer les patients atteints de la maladie et ceux qui n'en ont pas.

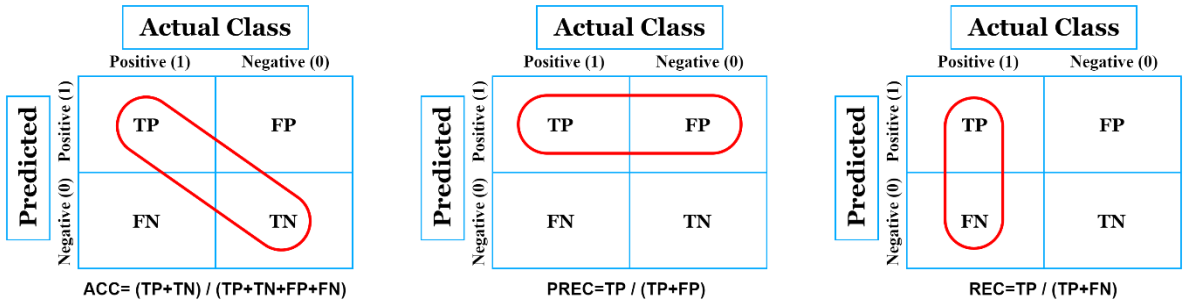

Figure III.7 - Mesures de performance : exactitude, précision, rappel.

Il existe d'autres mesures comme l'erreur quadratique moyenne (MSE), l'erreur quadratique moyenne racine (RMSE), l'erreur absolue moyenne (MAE), mais elles sont généralement utilisées dans les problèmes de régression. Par conséquent, cette étude comparative s'appuiera sur les mesures de performance expliquées ci-dessus en raison de l'ensemble de données et des algorithmes utilisés classés dans les problèmes de classification. De plus, les mêmes métriques sont utilisées pour évaluer la qualité de notre modèle d'ontologie.

Dans la section suivante, nous partagerons et résumerons le résultat obtenu à partir des classificateurs à l'aide des logiciels Weka et Protégé.

III.5. Résultats et discussion

Nous présentons dans cette section les résultats de l'évaluation des classificateurs employés dans cette étude, y compris le résultat et les statistiques du classificateur d'ontologie. Les résultats du classificateur d'ontologie sont fournis dans la Table III.3, Table III.4 et la Figure III.8 en utilisant les métriques de performance décrites dans la section précédente. De plus, nous fournissons les résultats de l'exactitude, de la précision, du rappel et de la F-mesure dans les figures (Figure III.9, Figure III.10, Figure III.11, Figure III.12), qui illustrent le visuel de chaque métrique. La Table III.5 décrit plus en détail les résultats expérimentaux pour les classificateurs d'apprentissage automatique et d'ontologie qui ont été utilisés dans cette recherche.

Table III.3 - Classificateur d'ontologie basé sur le mode de validation croisée 10 fois.

Matrice de confusion		**Classe réelle**	
		positive	**négative**
Classe prédite	**positive**	TP : 35525	FP : 9502
	négative	FN : 7660	TN : 17289

Table III.4 - Classificateur d'ontologie basé sur une validation en mode fractionné à 60 %.

Matrice de confusion		**Classe réelle**	
		positive	**négative**
Classe prédite	**positive**	TP : 35681	FP : 9295
	négative	FN : 7720	TN : 17280

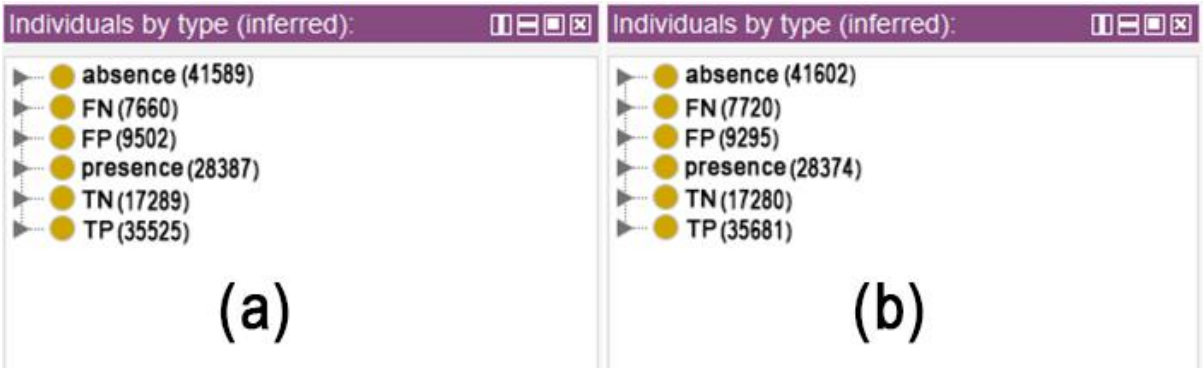

Figure III.8 - Résultats des concepts inférés. (a) validation croisée 10 fois. (b) Validation en mode fractionné à 60 %.

- **Exactitude :**

Dans la Figure III.9 et la Table III.5, nous avons obtenu la valeur la plus élevée en termes de mode de validation croisée 10 fois pour l'ontologie, l'arbre de décision et la régression logistique avec 75,5 %, 73,1 %, 72,1 % en conséquence. Presque les mêmes résultats en utilisant le mode de test fractionné, nous avons obtenu 75,7 %, 73,1 % et 72,3 % pour l'ontologie, l'arbre de décision et la régression logistique consécutivement.

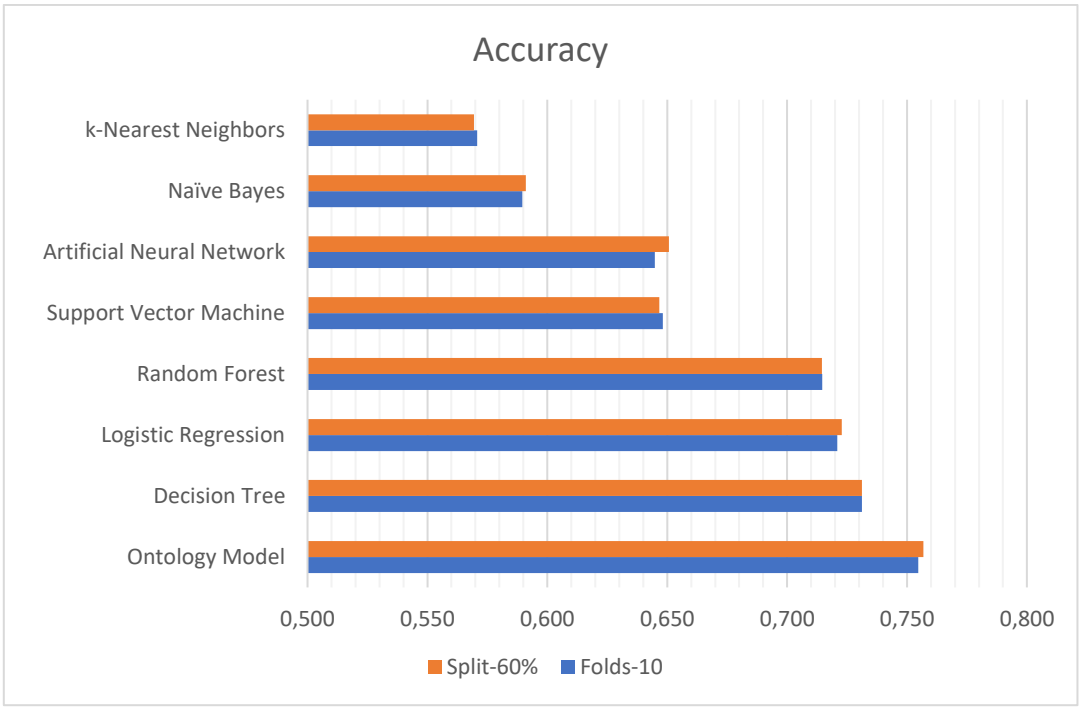

Figure III.9 - Résultats de comparaison de l'exactitude.

- **Précision :**

Le classificateur d'ontologie a la précision la plus élevée de 78,9 % et 79,3 % pour les deux modes de test. Suivi par Arbre de décision et Forêt aléatoire. Plus de détails sont présentés dans la Table III.5 et la Figure III.10.

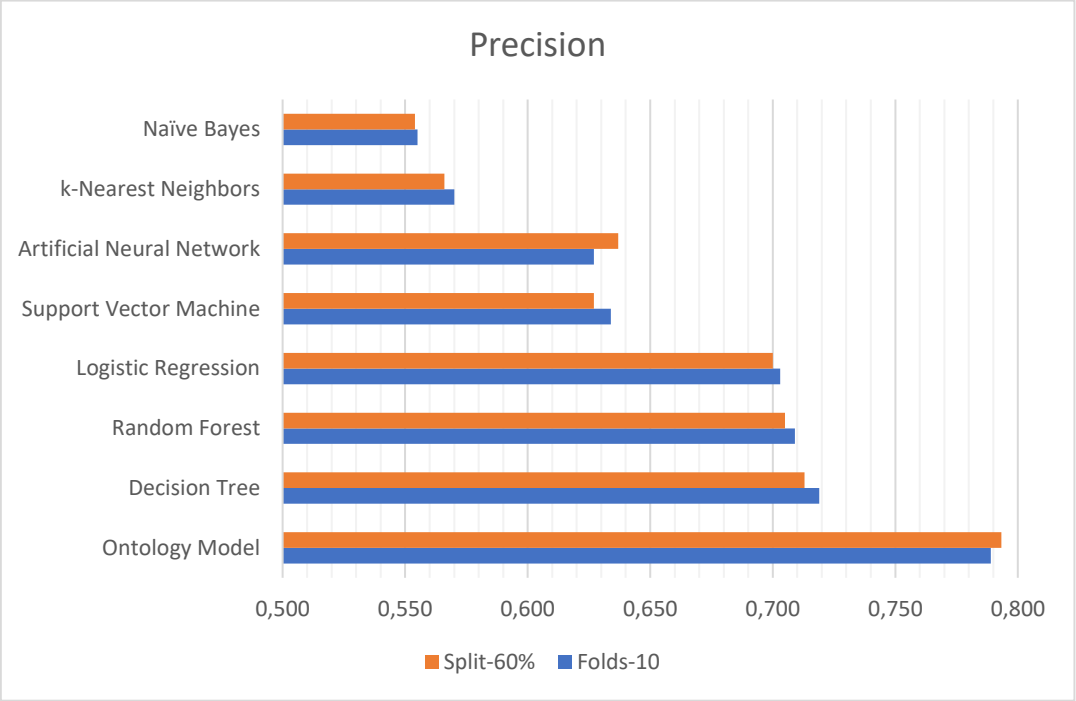

Figure III.10 - Comparaison des résultats de précision.

- **Rappel :**

D'après la Figure III.11 et la Table III.5, nous remarquons que Naïve Bayes avait la valeur la plus élevée dans les deux modes de test, suivi par l'Ontologie en deuxième position et Régression logistique avec Arbre de décision en troisième position.

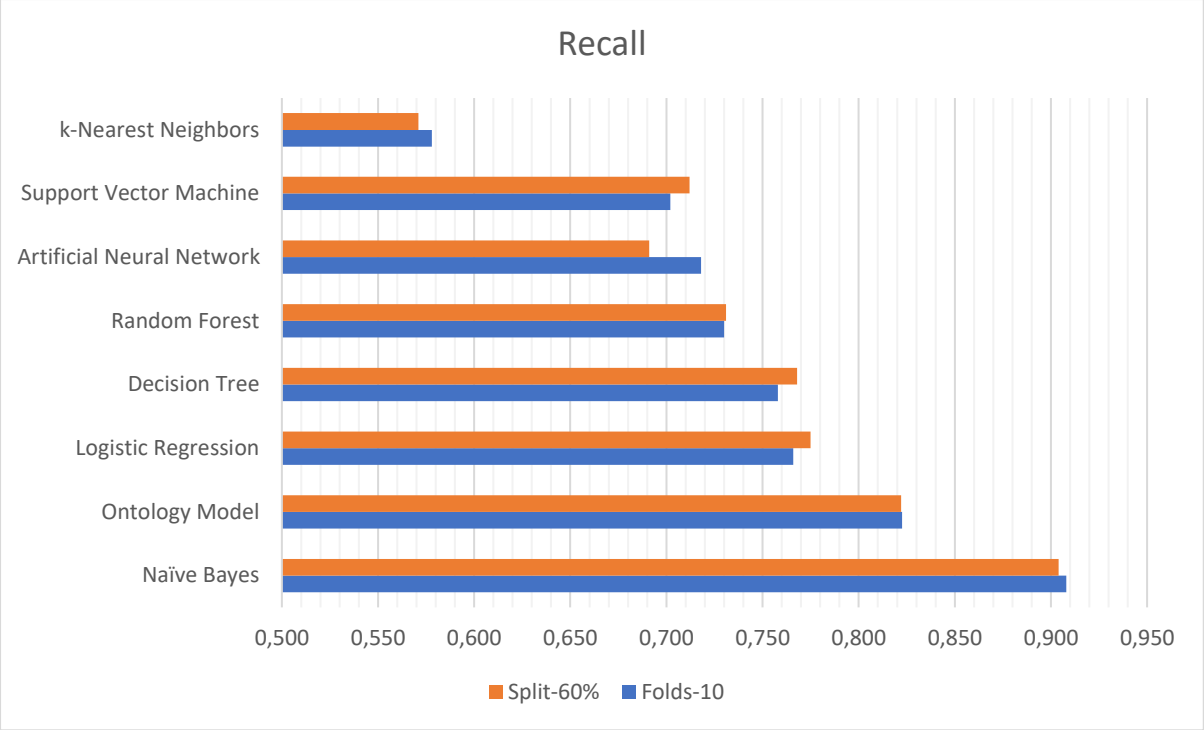

Figure III.11 - Comparaison des résultats du rappel.

- **F-Mesure :**

D'après la Figure III.12 et la Table III.5, nous remarquons que le modèle d'ontologie avait la valeur la plus élevée dans les deux modes de test, suivi de l'arbre de décision avec régression logistique en deuxième position et de la forêt aléatoire en troisième position.

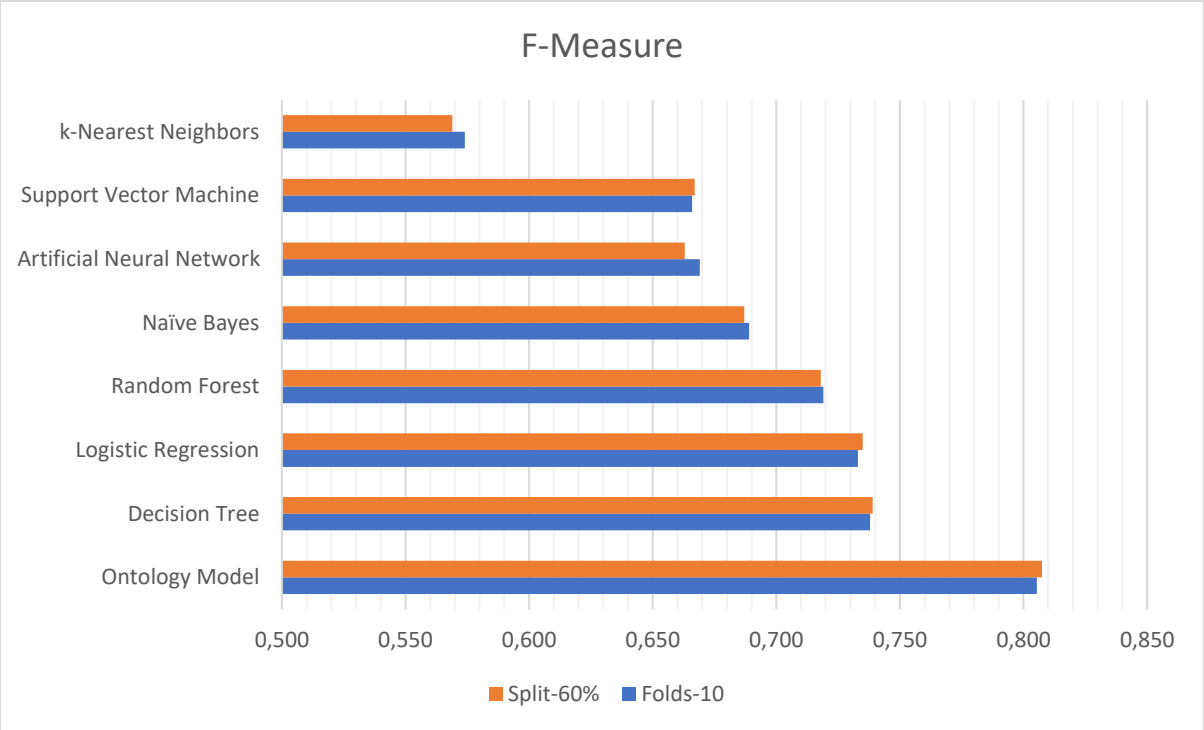

Figure III.12 - Résultats de comparaison de F-Mesure.

En ce qui concerne les résultats discutés ci-dessus, nous remarquons qu'il n'y a pas de grande différence entre la validation croisée et le mode de test fractionné en pourcentage. Les résultats expérimentaux montrent que le classificateur d'ontologie est considéré comme le meilleur avec une précision élevée de 75,5 %, suivi de l'arbre de décision de 73,1 % et de la régression logistique de 72,1 %. Nous concluons que la combinaison de l'apprentissage automatique avec le raisonnement ontologique (c'est-à-dire l'extraction de règles à partir d'algorithmes d'apprentissage automatique et leur intégration dans l'ontologie à l'aide de SWRL) peut fournir de meilleurs résultats. De plus, ces résultats comparatifs démontrent comment la représentation des connaissances et les capacités de raisonnement de l'ontologie OWL pourraient apporter des avantages supplémentaires en plus de la classification. De plus, comme le classificateur d'ontologie est un modèle interprétable, il peut offrir des informations sur la manière dont le processus parvient à la décision. Le classificateur d'ontologie produit des résultats équivalents et comparables aux classificateurs d'apprentissage automatique. Les résultats peuvent également être interprétés par des humains, et les règles peuvent être modifiées ou ajoutées au besoin.

À notre connaissance, il s'agit de la première analyse comparative de l'apprentissage automatique et des classificateurs d'ontologie, dans laquelle nous avons intégré l'ontologie à l'apprentissage automatique et plus particulièrement dans le domaine de la prédiction des maladies cardiovasculaires. Ainsi, aucune comparaison significative ne peut être faite pour cette raison, d'une part, d'autre part, les chercheurs utilisent différents ensembles de données et différentes méthodes de sélection et d'amélioration des performances.

Table III.5 - Résultats des classificateurs de l'apprentissage automatique et de l'ontologie.

	Exactitude		**Précision**		**Rappel**		**F-mesure**	
	Folds-10	Split-60%	Folds-10	Split-60%	Folds-10	Split-60%	Folds-10	Split-60%
K-Nearest Neighbors	0,571	0,569	0,57	0,566	0,578	0,571	0,574	0,569
Naïve Bayes	0,590	0,591	0,555	0,554	0,908	0,904	0,689	0,687
Artificial Neural Network	0,645	0,651	0,627	0,637	0,718	0,691	0,669	0,663
Support Vector Machine	0,648	0,647	0,634	0,627	0,702	0,712	0,666	0,667
Random Forest	0,715	0,715	0,709	0,705	0,73	0,731	0,719	0,718
Logistic Regression	0,721	0,723	0,703	0,7	0,766	0,775	0,733	0,735
Decision Tree	0,731	0,731	0,719	0,713	0,758	0,768	0,738	0,739
Ontology Model	0,755	0,757	0,789	0,793	0,823	0,822	0,805	0,807

III.6. Conclusion

Les techniques d'apprentissage automatique sont largement utilisées dans toutes les disciplines scientifiques et ont révolutionné les industries du monde entier. L'application d'outils et d'algorithmes d'apprentissage automatique dans les soins de santé a récemment connu des progrès significatifs [241], [242]. Ces procédés ont démontré leur efficacité et peuvent être bénéfiques dans le traitement de maladies chroniques telles que les maladies cardiovasculaires. De plus, le Web sémantique, pour sa part, a démontré sa valeur et sa force dans diverses disciplines, dont la santé. L'ontologie, en tant que composant du Web sémantique, a la capacité de traiter les concepts et les relations de la même manière que les humains voient les concepts connectés.

Dans ce chapitre, nous avons présenté sept algorithmes d'apprentissage automatique et un modèle d'ontologie, et nous avons expliqué leur évaluation comparative. De plus, différentes métriques de performance sont utilisées pour évaluer les résultats tels que l'exactitude, la précision, le rappel, la F-mesure.

Les résultats révèlent que, même sans sélection de caractéristiques appliquée, la méthode de classification ontologique a la plus grande précision. Cela nous amène à un nouveau champ de recherche que nous suggérons et encourageons les chercheurs à contribuer et à créer de nouvelles idées dans le même contexte, pour donner plus de résultats et de comparaison, à des fins de prédiction, de recommandation ou de prise de décision, etc. De notre côté, nous sommes impatients d'améliorer cette étude comparative en appliquant de nouvelles approches pour intégrer les règles de l'apprentissage automatique à la méthode de classification des ontologies, ainsi qu'en utilisant des algorithmes d'apprentissage automatique par régression.

CHAPITRE IV - INTEGRATION DE L'ONTOLOGIE AVEC L'APPRENTISSAGE AUTOMATIQUE POUR PREDIRE LA PRESENCE DE COVID-19 EN FONCTION DES SYMPTOMES

IV.1. Introduction

Le coronavirus (COVID-19) est apparu en 2019, appelé coronavirus du syndrome respiratoire aigu sévère 2 (SARS-COV-2). L'organisation mondiale de la santé (OMS) a déclaré que le COVID-19 était une pandémie mondiale en mars 2020, et les données ont confirmé que le COVID-19 se transmet d'une personne à une autre par mélange et proximité entre les personnes à une distance d'environ deux mètres. Quant à la méthode de propagation, c'est par les gouttelettes respiratoires, lorsqu'une personne infectée par le virus éternue, tousse ou respire et qu'une autre personne proche de lui l'inhale ou entre dans sa bouche, son nez ou ses yeux. Le coronavirus peut également être transmis par une personne infectée mais ne présentant pas de symptômes, soit par voie aérienne, soit en touchant une surface recouverte du virus puis en touchant directement la bouche, le nez ou les yeux de la personne. Les symptômes du virus covid-19 apparaissent dans les 14 jours suivant l'exposition au virus et comprennent de la fièvre, de la toux, une perte d'odorat ou de goût, un essoufflement, des douleurs musculaires, des nausées, de la diarrhée, des maux de gorge, des douleurs thoraciques et des frissons. La “Food and Drug Administration” des États-Unis s'est appuyée sur l'utilisation du vaccin pour prévenir l'infection par le virus COVID-19 , elle s'est donc appuyée sur l'utilisation de l'antiviral Pfizer-Biontech pour les personnes âgées de 15 ans et plus, et son utilisation dans les situations d'urgence pour les enfants âgés 5 à 15 ans, et pour les personnes âgées de 18 ans Plus que cela, la “Food and Drug Administration” des États-Unis s'est appuyée sur l'utilisation du vaccin Moderna pour les empêcher d'être infectés par le virus COVID-19. Quant aux personnes atteintes de maladies chroniques, elles devraient consulter un médecin sur d'autres moyens de se protéger contre l'infection par le COVID-19. Pour éviter l'infection par le COVID-19, l'OMS et le centre de contrôle et de prévention des maladies ont recommandé que certaines mesures de précaution soient prises, notamment les suivantes :

- Évitez les rassemblements dans les endroits bondés et fermés et gardez une distance de deux mètres entre vous et les autres.
- Recevez le vaccin.
- Se laver les mains avec de l'eau et du savon pendant 20 secondes ou utiliser un désinfectant pour les mains à base d'alcool avec une concentration d'au moins 60 %.
- Portez un masque dans les lieux publics fermés.
- Nettoyer et désinfecter les surfaces.
- Lorsque vous toussez ou éternuez, évitez de vous toucher les yeux, le nez ou la bouche et couvrez le nez et la bouche avec un mouchoir ou un coude.

Prédire le risque de gravité de toute maladie à un stade précoce est une tâche cruciale et a de nombreux effets, comme la réduction du taux de mortalité, la consommation de ressources hospitalières et le soutien aux médecins dans leur prise de décision. Durant la période critique, pendant la propagation du coronavirus dans le monde et le nombre croissant de

patients et de décès, le nombre de patients COVID-19 a atteint près de 230 millions alors que le nombre de décès était de 4,7 millions dans le monde jusqu'à présent lors de la rédaction de cette recherche, selon les statistiques de l'Université Johns Hopkins [243]. Les États-Unis sont à la tête des pays, suivis du Brésil, de l'Inde, de la France, de la Russie, de l'Italie et de nombreux autres pays. Les raisons de cette croissance en nombre sont la forte prévalence de COVID-19, un diagnostic tardif et le manque de ressources dans de nombreux hôpitaux pour absorber cette pandémie. Par conséquent, prédire le risque de gravité des patients COVID-19 est une tâche critique et a de nombreux résultats positifs, tels que la fourniture des soins de santé requis pour chaque patient en fonction de sa gravité, une bonne consommation des ressources hospitalières qui accordent la plus haute priorité au haut patient à risque, et d'aider les médecins à prendre leurs décisions qui conduiront à l'amélioration du traitement du patient.

Dans ce chapitre, nous avons l'intention de comparer sept approches de classification importantes avec le modèle ontologique en utilisant des critères soigneusement choisis obtenus à partir de la matrice de confusion, notamment la F-mesure, l'exactitude, le rappel et la précision. Le reste de ce chapitre est organisé comme suit : la section 2 décrit les méthodologies utilisées dans cette analyse comparative. La section 3 résume les résultats et la discussion.

IV.2. Revue de littérature

Trois ressources principales qui pourraient être utilisées pour détecter le COVID-19 : les images radiographiques, la tomodensitométrie (CT) et la réaction en chaîne par polymérase de transcription inverse (RT-PCR). Le meilleur type est la RT-PCR, mais elle est très coûteuse, n'est pas disponible dans tous les hôpitaux et prend beaucoup de temps pour obtenir les résultats. Par conséquent, de nombreux médecins dépendent de l'imagerie radiologique thoracique telle que les rayons X et la tomodensitométrie pour le diagnostic précoce et le traitement de cette maladie [244]. La tomodensitométrie est un outil très sensible, mais ses résultats peuvent être observés après une longue période en fonction de l'apparition des symptômes, où la tomodensitométrie normale prend de zéro à deux jours pour voir ses résultats [245], de sorte que la tomodensitométrie est difficile à utiliser pour surveiller les patients périodiquement. La radiographie thoracique (CXR) est moins sensible que la TDM et la RT-PCR, mais c'est l'une des méthodes les plus couramment utilisées et les plus accessibles pour l'examen rapide des affections pulmonaires. Les résultats des rayons X sont observés en peu de temps et ce n'est pas une technique coûteuse, elle peut donc être utilisée périodiquement pour surveiller l'état du patient.

L'apprentissage automatique s'est révélé être un domaine d'étude de premier plan au cours de la dernière décennie en résolvant de nombreux problèmes du monde réel très complexes et sophistiqués [246]. Les domaines d'application comprenaient presque tous les domaines du monde réel tels que les soins de santé, les véhicules autonomes, les applications commerciales, le traitement du langage naturel, les robots intelligents, les jeux, la modélisation climatique, le traitement de la voix et des images. L'apprentissage des algorithmes d'apprentissage automatique est généralement basé sur une méthode d'essais et d'erreurs tout à fait opposée aux algorithmes conventionnels, qui suit les instructions de

programmation basées sur des déclarations de décision comme si-sinon [247]. L'un des domaines les plus importants de l'apprentissage automatique est la prévision [248], de nombreux algorithmes d'apprentissage automatique standard ont été utilisés dans ce domaine pour guider le futur plan d'action nécessaire dans de nombreux domaines d'application, notamment les prévisions météorologiques, les prévisions de maladies, les prévisions boursières ainsi comme pronostic de la maladie. Divers modèles de régression et de réseau neuronal ont une large applicabilité pour prédire les conditions des patients à l'avenir avec une maladie spécifique [249]. De nombreuses études ont été réalisées pour la prédiction de différentes maladies à l'aide de techniques d'apprentissage automatique telles que la maladie coronarienne [250], la prédiction des maladies cardiovasculaires [251], [252] et la prédiction du cancer du sein [19], [253]. En particulier, l'étude [254] se concentre sur la prévision en direct des cas confirmés de COVID-19 et l'étude [255] se concentre également sur la prévision de l'épidémie de COVID-19. Ces systèmes de prédiction peuvent être très utiles dans la prise de décision pour gérer le scénario actuel afin de guider les interventions précoces pour gérer ces maladies très efficacement.

Récemment, les chercheurs ont publié une quantité importante de recherches utilisant des algorithmes d'apprentissage automatique pour diagnostiquer le covid-19 [256]–[258]. Dans cette analyse comparative [259], les auteurs visent à déterminer quelle technique de classification a le taux de précision le plus élevé pour les échantillons de données positifs au covid-19 collectés, les résultats donnent 85 %, 80 % et 65 % de précision, pour la machine à vecteurs de support, K-Nearest Neighbors et Naive Bayes respectivement. Les auteurs d'une autre étude [260] ont mené une analyse basée sur des incidents survenus dans différents états de l'Inde dans l'ordre chronologique. Ils ont effectué le nettoyage des données et la sélection des caractéristiques sur l'ensemble de données, suivis de la prévision de toutes les classes à l'aide d'un réseau de neurones, d'une machine à vecteurs de support, d'un modèle linéaire, d'une forêt aléatoire et d'un arbre de décision, où le modèle de forêt aléatoire a surpassé les autres.

L’étude [261] visait à construire un modèle prédicteur de présence COVID-19 en appliquant cinq algorithmes d'apprentissage automatique supervisé, dont J48 Decision Tree, Random Forest, K-Nearest Neighbors, Naïve Bayes et Support Vector Machine. Une analyse comparative a été menée en évaluant les performances du modèle en validation croisée 10 fois via le logiciel d'apprentissage automatique WEKA. Les résultats montrent que la machine à vecteurs de support utilisant le noyau universel Pearson VII est le meilleur algorithme d'apprentissage automatique, avec une précision de 98,81 % et une erreur absolue moyenne de 0,012. L'algorithme Support Vector Machine a surpassé les autres algorithmes en termes d'exactitude, de précision, de rappel, de F-mesure, d'instances classées correctement et incorrectement, de score statistique kappa, d'erreur absolue moyenne et de temps nécessaire à la construction du modèle. De plus, les résultats montrent également que Random Forest est le deuxième meilleur algorithme à prendre en compte dans la construction d'un prédicteur de présence COVID-19, car il a les mêmes mesures de précision que celles obtenues par l'algorithme Support Vector Machine, à l'exception de l'erreur absolue moyenne. L'algorithme Random Forest peut être envisagé pour développer un modèle très performant avec un temps d'apprentissage plus rapide par rapport à la machine à vecteurs de support. De plus, K-Nearest Neighbors est le troisième algorithme le plus approprié à utiliser en termes de mesures de précision, car il peut également construire un modèle en peu de temps par rapport

à d'autres algorithmes. Ensuite, l'arbre de décision J48 est classé au quatrième rang et Naïve Bayes est classé au cinquième rang des algorithmes les plus appropriés à prendre en compte.

Dans cette étude [262], Une approche de classification d'ensemble affinée pour prédire les taux de décès et de guérison des patients infectés à l'aide de techniques d'apprentissage automatique a été proposée pour différents états de l'Inde. Le modèle de classification proposé est appliqué au récent ensemble de données COVID-19 pour l'Inde, et une évaluation des performances de divers classificateurs de pointe du modèle proposé est effectuée. Les classificateurs ont prévu le statut infectieux des patients dans différentes régions afin de mieux planifier les ressources et les systèmes de soins de réponse. La classification appropriée de la classe de sortie basée sur les entités d'entrée extraites est essentielle pour obtenir des résultats précis des classificateurs. Le résultat expérimental montre que le modèle hybride proposé a atteint un F1-score maximal de 94 % par rapport aux ensembles et à d'autres classificateurs tels que Support Vector Machine, Decision Tree et Gaussian Naïve Bayes sur un ensemble de données de 5004 instances via une validation croisée de 10 fois pour prédire la bonne classe. La faisabilité de la prédiction automatisée des taux de guérison et de mortalité de l'infection au COVID-19 dans les États indiens a été démontrée.

L'article [263] présente une étude détaillée des modèles de prévision récemment développés et prédit le nombre de cas confirmés, récupérés et mortels en Inde causés par le COVID-19. Les coefficients de corrélation et la régression linéaire multiple appliqués pour la prédiction et l'autocorrélation et l'autorégression ont été utilisés pour améliorer la précision. Le nombre prédit de cas montre un bon accord avec un score de 0,9992 R-carré aux valeurs réelles. La découverte suggère que le verrouillage et la distanciation sociale sont deux facteurs importants qui peuvent aider à supprimer le taux de propagation croissant de COVID-19.

L'objectif de ce travail [264] est d'explorer l'apprentissage automatique et de développer un modèle COVID-19 capable de prédire le nombre de cas avec une grande précision. L'étude proposée utilise des modèles SVR et PR pour prévoir le nombre de cas récupérés, de cas confirmés, de décès et le nombre de cas quotidiens. Les données sont collectées du 1er mars au 30 avril 2020. Le nombre confirmé de cas au 30 avril était de 35043, avec 1147 décès au total et 8889 patients guéris. Le modèle a été créé en Python. Ils ont examiné divers algorithmes de prédiction d'apprentissage automatique. En conclusion, les algorithmes d'apprentissage supervisés se sont avérés meilleurs que les algorithmes d'apprentissage non supervisés. Ces modèles de prédiction peuvent nous aider à nous préparer à une autre vague de COVID-19 et à garantir la disponibilité des ressources nécessaires.

Cette étude [265] visait à comparer plusieurs algorithmes d'apprentissage automatique pour prédire la mortalité par COVID-19 en utilisant les données du patient lors de la première admission et choisir l'algorithme le plus performant comme outil prédictif pour la prise de décision. Après la sélection des caractéristiques, basée sur les prédicteurs confirmés, des informations sur 1500 patients éligibles (1386 survivants et 144 décès) obtenues à partir du registre de l'hôpital Ayatollah Taleghani, ville d'Abadan, Iran, ont été extraites. Par la suite, plusieurs algorithmes d'apprentissage automatique ont été formés pour prédire la mortalité par COVID-19. Finalement, pour évaluer les performances des modèles, les métriques issues de la matrice de confusion ont été calculées. Les participants à l'étude étaient 1500 patients ; le nombre d'hommes est supérieur à celui des femmes (836 contre 664) et l'âge moyen est de

57,25 ans (interquartile 18-100). Après avoir effectué la sélection des caractéristiques, sur 38 caractéristiques, la dyspnée, l'admission aux soins intensifs et l'oxygénothérapie ont été identifiées comme les trois principaux prédicteurs. Le tabagisme, l'alanine aminotransférase et la numération plaquettaire se sont révélés être les trois facteurs prédictifs les plus faibles de la mortalité par COVID-19. Les résultats expérimentaux ont démontré que la forêt aléatoire (RF) avait de meilleures performances que les autres algorithmes d'apprentissage automatique avec une exactitude, une sensibilité, une précision, une spécificité et une caractéristique de fonctionnement du récepteur (ROC) de 95,03 %, 90,70 %, 94,23 %, 95,10 % et 99,02 %, respectivement. Il a été constaté que l'apprentissage automatique permet un niveau de précision raisonnable dans la prédiction de la mortalité par COVID-19. Par conséquent, les modèles prédictifs basés sur l'apprentissage automatique, en particulier l'algorithme de forêt aléatoire, facilitent potentiellement l'identification des patients à haut risque de mortalité et informent les interventions appropriées des cliniciens.

Une analyse des ensembles de données COVID-19 pour comprendre quel groupe d'âge est principalement affecté par le COVID-19 est menée dans cette étude [266]. Différents modèles de prédiction sont construits à l'aide d'algorithmes d'apprentissage automatique et leurs performances sont calculées et évaluées. Random Forest Regressor et Random Forest Classifier ont surpassé les autres modèles d'apprentissage automatique comme Support Vector Machine, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, Multilinear Regression, Logistic Regression et XGBoost Classifier. Les expériences révèlent que les personnes des tranches d'âge 20-30, 30-40 et 40-50 souffrent de COVID-19. Les matrices de corrélation sont construites pour comprendre la relation entre les caractéristiques des ensembles de données. L'importance des caractéristiques est calculée pour les classificateurs construits. Avec les classificateurs et les répresseurs sont également construits pour la prédiction. Les résultats montrent que le Random Forest Regressor et le Random Forest Classifier ont surpassé les autres modèles en termes de CoD (Coefficient of

Determination) et de Précision.

Cette étude [267] a suggéré un Framework qui a détecté les personnes infectées par COVID-19 avec une grande précision. La maladie COVID-19 est contagieuse, par conséquent, la surveillance à distance avec la technologie basée sur IoMT est la meilleure solution pour le contrôler. Comme les cas ont augmenté à un rythme exponentiel, il est nécessaire d'identifier chaque cas positif pendant cette urgence. Dans cette étude, un Framework d'apprentissage profond d'ensemble est proposé qui utilise les précisions prédictives des modèles d'apprentissage par transfert individuels, à savoir (i) ResNet50, (ii) DenseNet201, (iii) InceptionV3, (iv) VGG-16, (v) VGG-19, (vi) Xception, et (vii) MobineNetV2 pour prédire les personnes infectées par le COVID-19. Le Framework d'apprentissage d'ensemble utilise la force individuelle des apprenants transférés pour détecter le COVID-19 à partir des images de radiographie pulmonaire. Malgré ses lourdes exigences de calcul et sa structure complexe, ce Framework est suffisamment pratique car il fournit également des résultats optimaux sur l'ensemble de données de validation. Cependant, le système proposé a des limites. Tout d'abord, le jeu de données est assez petit pour tester la généralisation du système. Cela peut être résolu si plus d'images sont utilisées pour entraîner le modèle. Deuxièmement, à partir de maintenant, le modèle fonctionne sur la vue postéro-antérieure (PA) des rayons X. Par conséquent, il ne peut pas différencier les vues antéro-postérieures (AP), latérales, etc. Troisièmement, le COVID-19 à des degrés divers ne peut pas être identifié pour l'instant.

Dans cette étude [268], les auteurs ont démontré une prédiction très précise des « all-cause mortality/cardiac arrest » (AM/CA) et « imaging-confirmed thromboembolic events » TEs chez les patients hospitalisés COVID-19 en utilisant le prédicteur COVID-HEART mis à jour en continu. Dans sa mise en œuvre actuelle, le prédicteur peut faciliter des changements pratiques et significatifs dans le triage des patients et l'allocation des ressources en fournissant des scores de risque en temps réel pour les complications survenant couramment chez les patients COVID-19. Le COVID-HEART peut être recyclé pour prédire d'autres événements CV indésirables, notamment l'infarctus du myocarde et l'arythmie. L'utilité potentielle du prédicteur s'étend bien au-delà des patients hospitalisés COVID-19, car COVID-HEART pourrait être appliqué à la prédiction des événements indésirables après la sortie de l'hôpital ou chez les patients atteints du syndrome COVID chronique ("long COVID"). De plus, la méthodologie ML utilisée ici pourrait être étendue pour être utilisée dans d'autres scénarios cliniques qui nécessitent un dépistage ou une détection précoce, tels que le risque de réadmission à l'hôpital, dans le but d'améliorer les résultats cliniques grâce à des avertissements précoces et à la possibilité d'une intervention rapide.

Actuellement, les images radiographiques sont utilisées comme premiers symptômes pour détecter les patients COVID-19. Par conséquent, une recherche qui été effectuée [269], dans cette recherche un modèle de prédiction a été construit pour prédire différents niveaux de risques de gravité pour le patient COVID-19 sur la base d'images radiographiques en appliquant des techniques d'apprentissage automatique. Pour construire le modèle proposé, un modèle pré-formé profond CheXNet et des techniques artisanales hybrides ont été appliqués pour extraire les caractéristiques, deux méthodes différentes : l'analyse en composantes principales (PCA) et l'élimination récursive des caractéristiques (RFE) ont été intégrées pour sélectionner les caractéristiques les plus importantes, puis, six techniques d'apprentissage automatique ont été appliquées. Pour les fonctionnalités artisanales, les expériences ont prouvé que la fusion des fonctionnalités qui ont été sélectionnées par PCA et RFE ensemble (PCA + RFE) a obtenu les meilleurs résultats avec tous les classificateurs par rapport à l'utilisation de toutes les fonctionnalités ou à l'utilisation des fonctionnalités sélectionnées par PCA ou RFE individuellement. Le classificateur XGBoost a obtenu les meilleures performances avec les fonctionnalités fusionnées (PCA + RFE), où il a atteint 97 % de précision, 98 % de précision, 95 % de rappel, 96 % de score f1 et 100 % de roc-auc. De plus, SVM a obtenu les mêmes résultats avec quelques différences mineures, mais dans l'ensemble, c'était une bonne performance où il a atteint 97% de précision, 96% de précision, 95% de rappel, 95% de F1-score et 99% de roc-auc. D'autre part, pour les fonctionnalités CheXNet pré-formées, les classificateurs Extra Tree et SVM avec RFE ont atteint 99,6 % pour toutes les mesures.

Dans cette étude [270], la propagation du COVID-19 dans différents États indiens est discutée, et un modèle d'ensemble utilisant la régression linéaire, la régression polynomiale, la régression SVM est proposé et vérifié expérimentalement pour prévoir les cas confirmés de COVID-19 en Inde. Tous les modèles sont évalués pour leur exactitude dans cette étude. Lorsque les modèles sont comparés, ils ont observé que le modèle d'ensemble fournit des valeurs prédites plus précises pour la prévision des données de séries chronologiques que les autres modèles. Selon la découverte, beaucoup plus d'exigences de restriction COVID-19 sont nécessaires pour contrôler la propagation de la maladie. La prédiction pourrait aider à la

prise de décision en matière de soins de santé et des mesures proactives pourraient être prises pour réduire les pertes en vies humaines. Le modèle d'ensemble proposé peut être étendu pour la prédiction de la récupération et des décès à un certain endroit.

Dans cette étude [271], les auteurs ont appliqué différentes techniques d'apprentissage automatique telles que la forêt aléatoire, l'arbre de décision, la régression linéaire, la recherche binaire et le k-plus proche voisin sur l'ensemble de données des patients mexicains pour découvrir l'impact des maladies à vie sur l'augmentation des symptômes du virus dans le corps humain. Ils ont obtenu des résultats suffisants en classant les cas des patients en fonction de leurs maladies à vie. Le système proposé utilisant cinq algorithmes d'apprentissage automatique (RF, DT, SVM, KNN et LR) a classé l'ensemble de données avec une précision de (0.88 %, 0.88 %, 0.87, 0.86 et 0.88 %) respectivement.

Pour réduire la propagation du COVID-19 et prendre les bonnes décisions pour le gouvernement, des prévisions de cas futurs sont nécessaires. Par conséquent, l'algorithme du prophète est utilisé pour prédire la propagation du COVID-19 pour l'année à venir [272], les auteurs de cette étude ont choisi l'algorithme du prophète car cet algorithme a un niveau de précision assez élevé. Pour les auteurs, cette recherche visait à prouver l'exactitude du modèle de prédiction sur une période d'un an (du 2 mars 2020 au 12 février 2021). Le pourcentage de patients confirmés était de 22,60 à 42,11 %, les décès de 21,67 à 39,00 % et les guérisons de 22,53 à 41,82 %. Le pourcentage prévu de cas décédés était de 2,43 % chaque jour. Ensuite, pour le niveau de précision, les auteurs utilisent la prédiction et les données originales du 13 février 2021 au 3 mai 2021. Sur la base des résultats de l'exactitude, c'est remarqué que les données des patients atteints de covid-19 confirmé obtiennent une valeur inférieure à 50 %, ce qui est différent des données sur la mort et la guérison qui atteignent un taux d'exactitude de plus de 50 %. Cela peut être dû au processus de modélisation où le modèle de prédiction n'était pas si proche du modèle à partir des données d'origine. Cela signifie que l'utilisation de l'algorithme Prophet pour prédire la propagation du COVID-19 en Indonésie n'est toujours pas suffisamment valide, donc une validation ou une comparaison est nécessaire à l'aide d'autres algorithmes logistiques.

Dans le domaine médical, les ontologies ont été appliquées non seulement pour la modélisation et l'analyse des données mais également pour la classification des données. En effet, elles sont liées plusieurs techniques de classification à savoir l'apprentissage automatique ou le deep learning (apprentissage profond). Le fait de combiner les ontologies aux techniques issues des domaines de l'apprentissage automatique caractérise un domaine récent permettant de concevoir de nouvelles stratégies de raisonnement. De plus, l'ontologie a été l'une des techniques les plus largement utilisées pour gérer, organiser et extraire des données au cours des dernières décennies. C'est un mode de représentation des données qui a été efficacement utilisé dans un certain nombre de domaines, en particulier le domaine médical. Il est important en informatique en raison de sa capacité à exprimer de nombreux concepts et leurs relations entre les domaines. En réalité, aucune ontologie unique n'est suffisante pour répondre aux demandes croissantes de soins de santé d'aujourd'hui, et les ontologies doivent être combinées avec des algorithmes d'apprentissage automatique pour faciliter l'intégration et l'analyse des données.

IV.3. Méthodes et évaluation

Les approches et les matériaux utilisés, ainsi que la méthodologie expérimentale, la description de l'ensemble de données, les algorithmes d'apprentissage automatique, le modèle d'ontologie et les mesures d'évaluation, sont tous inclus dans cette section. La Figure IV.1 illustre l'organigramme du processus de cette étude comparative.

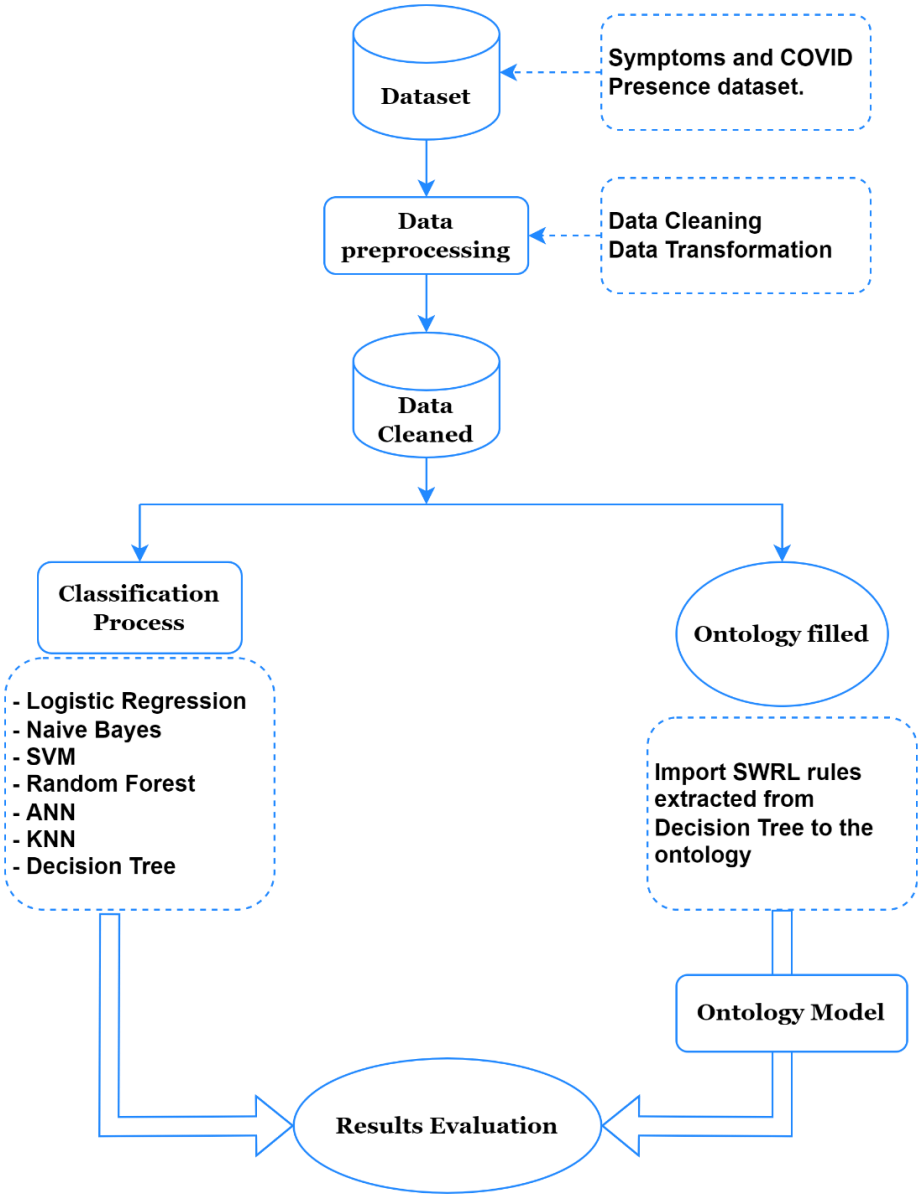

Figure IV.1 - Étapes méthodologiques.

IV.3.1. Prétraitement des données

Un ensemble de données accessible au public utilisé est nommé Symptoms and COVID-19 Presence[12] à partir du site Web Kaggle. L'ensemble de données contient des données sur les patients COVID-19. Il se compose de 5434 instances et de 21 fonctionnalités (20 attributs et

[12] https://www.kaggle.com/hemanthhari/symptoms-and-covid-presence

le dernier est une cible). Une description complète de tous les attributs de l'ensemble de données est fournie dans la Table IV.1.

Table IV.1 - Les attributs de l'ensemble de données

Attribut	*Description*
1- breath_pro	Problème respiratoire : la personne a du mal à respirer.
2- fever	La température est plus élevée que d'habitude.
3- dry_cough	Toux qui n'est pas accompagnée de mucosités.
4- sore_throat	L'individu souffre d'un mal de gorge.
5- run_nose	Nez qui coule : L'individu souffre d'un nez qui coule.
6- asthma	L'individu souffre d'asthme.
7- cld	Maladie pulmonaire chronique : L'individu souffre d'une maladie pulmonaire.
8- headache	L'individu souffre d'un mal de tête.
9- heart_disease	L'individu souffre d'une maladie cardiovasculaire.
10- diabetes	La personne est diabétique ou a des antécédents familiaux de diabète.
11- hyper_tension	Avoir une tension artérielle élevée.
12- fatigue	L'individu est fatigué.
13- gastrointestinal	L'individu a des problèmes gastriques.
14- abroad_travel	Récemment voyagé à l'extérieur du pays.
15- ccp	Contact avec un patient COVID : interaction avec une personne infectée par le COVID-19.
16- alg	Participation au Grand Rassemblement.
17- vpep	Visité des lieux publics exposés.
18- fwpep	Famille travaillant dans des lieux publics exposés.
19- wearing_masks	Porter des masques.
20- sm	Désinfection du marché : avant d'utiliser des objets achetés sur le marché, ils doivent être désinfectés.
21- COVID-19	Classe prédite (Présence ou absence de Coronavirus).

La phase de prétraitement des données vise à préparer les données à utiliser dans le modèle de prédiction. Habituellement, les données sont désordonnées et proviennent de différentes sources avec différentes tailles et résolutions. Cette phase est donc cruciale pour nettoyer et

normaliser les données afin de réduire la complexité et d'augmenter la précision du modèle de prédiction. Différents types de transformations peuvent être exécutés en fonction de l'ensemble de données, comme le redimensionnement, la rotation, le décalage, la normalisation, etc. Pour créer un classificateur d'apprentissage automatique efficace, nous devons toujours commencer par le nettoyage des données, la normalisation des fonctionnalités, la transformation des fonctionnalités et même la création de nouvelles fonctionnalités à partir de l'ensemble de données. L'ensemble de données que nous avons utilisé contient 4968 instances similaires, après suppression des instances en double, il reste 466 instances, où 385 représentent des individus avec covid-19 et 81 représentent des individus sans covid-19. Nous tenons à vous informer qu'afin de fournir une comparaison équitable des résultats de classification obtenus, nous n'avons utilisé aucune méthode de sélection de fonctionnalités ou d'amélioration des performances.

IV.3.2. Modèles de prédiction

Pour construire un modèle prédictif robuste, différents classificateurs d'apprentissage automatique ont été utilisés, parmi eux, nous citons : K plus proches voisins (KNN), forêt aléatoire (RF), régression logistique (LR), réseau de neurones artificiels (ANN), Naive Bayes (NB), arbre de décision (DT) et machine à vecteurs de support (SVM).

K Nearest Neighbors (KNN) est un algorithme supervisé développé par Thomas Cover [273] pour les problèmes de classification et de régression. Il utilise une méthode de similarité des caractéristiques pour prédire l'étiquette d'un nouveau point donné, ce qui signifie en outre que le nouveau point de test sera classé en fonction du vote majoritaire des K voisins les plus proches dans l'ensemble d'apprentissage, où K est le nombre de voisins. Il est caractérisé comme simple, facile à mettre en œuvre, ne dépendant que d'un seul paramètre (K), et de classificateurs efficaces.

Random Forest (RF) est un classificateur d'ensemble et une version améliorée du classificateur d'ensachage original. La forêt se compose de nombreux arbres de décision et le résultat final de la classification est déterminé en agrégeant tous les résultats de la classification de ces arbres composés et en prenant le vote majoritaire des résultats de la classification [274]. Il diffère de l'algorithme de bagging en utilisant un algorithme d'apprentissage qui sélectionne un sous-ensemble aléatoire des caractéristiques à chaque fractionnement de la phase de croissance. Pour la raison de rendre tous les arbres de décision différents, car chaque arbre utilise un sous-ensemble de données aléatoire différent.

Support vector machine (SVM) est un algorithme d'apprentissage automatique supervisé proposé par Cortes et Vapnik [275]. L'objectif de SVM est de trouver la frontière de décision optimale avec un hyperplan de marge maximale entre les différentes classes d'échantillons. Pour y parvenir, SVM doit convertir l'espace des données d'entrée d'un espace de faible dimension en un espace de dimension supérieure pour séparer les ensembles de données en différents échantillons avec la limite optimale. Cette conversion est implémentée par une technique appelée noyau. Le noyau convertit les problèmes non séparables en problèmes séparables en ajoutant plus de dimensions aux données. Les méthodes de noyau couramment utilisées comprennent le noyau à base radiale, le noyau polynomial et le noyau linéaire.

La régression logistique peut être utilisée pour effectuer une régression en tant que classification. Elle est basée sur la fonction prédictive sigmoïde définie par : $h(z) = \frac{1}{1+e^{-z}}$ où z est une fonction linéaire. La fonction renvoie un score de probabilité *P* compris entre 0 et 1. Afin de mapper cela à deux classes discrètes (0 ou 1), une valeur seuil θ est fixée. La classe prédite est égale à 1 si $P \geq \theta$, à 0 le cas contraire.

Un arbre de décision est un algorithme qui cherche à partitionner les individus en groupes d'individus aussi similaires que possible du point de vue de la variable à prédire. Le résultat de l'algorithme produit un arbre qui révèle les relations hiérarchiques entre les variables. Un processus itératif est utilisé où à chaque itération une sous-population d'individus est obtenue en choisissant la variable explicative qui permet la meilleure séparation des individus. L'algorithme s'arrête lorsqu'il n'y a plus de division possible.

Les réseaux de neurones artificiels sont un algorithme de classification supervisée populaire qui tente d'imiter le fonctionnement du cerveau humain. Il est souvent utilisé chaque fois qu'il existe de nombreuses données d'entraînement étiquetées avec de nombreuses fonctionnalités [276]. Le réseau calcule à partir de l'entrée un score (ou une probabilité) d'appartenir à chaque classe. La classe affectée à l'objet d'entrée correspond à celle avec le score le plus élevé. Un réseau de neurones est un système composé de neurones. Il est divisé en plusieurs couches connectées les unes aux autres où la sortie d'une couche correspond à l'entrée de la suivante [277]. Le calcul du score final est basé sur le calcul d'une fonction linéaire à partir des poids des couches et d'une fonction d'activation. Les valeurs de poids sont affectées aléatoirement à chaque entrée au début, puis sont apprises (mises à jour) par rétropropagation du gradient pour minimiser la fonction de perte associée à la couche finale. L'optimisation se fait avec une technique de descente de gradient [278].

L'algorithme K-Nearest Neighbor est l'un des algorithmes paresseux non paramétriques les plus simples et est couramment utilisé dans les techniques de classification avec diverses applications [279]. Le classificateur KNN mesure la similarité entre le nouveau point et chaque donnée d'apprentissage en recherchant le groupe de K voisins les plus identiques pour estimer la valeur ou la classe du nouveau point [280], [281]. Cela signifie qu'une valeur est attribuée au nouveau point en fonction de la proximité avec les points de l'ensemble d'apprentissage [282]. La distance de calcul est définie par différentes méthodes, dont les plus connues sont la distance euclidienne et la distance de Hamming où terme euclidien, la distance entre deux éléments, X= (x1, x2, ..., xn) et Y= (y1, y2, …, yn) est noté d (X, Y):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (yi - xi)^2}$$

Naïve Bayes est un simple algorithme d'apprentissage supervisé probabiliste dépendant du théorème de Bayes [73]. Il est réputé pour surpasser même les méthodes de classification très avancées. Il fournit un groupe de probabilités en calculant la fréquence des valeurs et leurs combinaisons dans un ensemble de données. Cependant, le classificateur Naïve Bayes est basé sur une hypothèse forte que les valeurs d'attribut sont indépendantes compte tenu de la valeur cible. Ainsi, il est qualifié de naïf car cette hypothèse indépendante n'est jamais vraie

dans les applications du monde réel. Le théorème de Bayes est une formule mathématique pour calculer les probabilités conditionnelles d'événements :

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

$p(A|B)$: Probabilité de A étant donné que B est vrai.

$p(A|B)$: Probabilité de B étant donné que A est vrai.

$p(A)$: Probabilité indépendante de A.

$p(B)$: Probabilité indépendante de B.

Nous avons choisi la méthode de l'arbre de décision parmi les sept classificateurs pour exécuter les règles produites par le modèle d'ontologie. L'algorithme d'arbre de décision est une technique d'apprentissage supervisé utilisée dans les statistiques, l'exploration de données et l'apprentissage automatique. Contrairement à d'autres algorithmes d'apprentissage supervisé, l'approche par arbre de décision peut également être utilisée pour résoudre des problèmes de régression et de classification. Nous avons choisi l'approche de l'arbre de décision pour plusieurs raisons, notamment le fait que la sortie de l'arbre de classification est plus facile à saisir et à analyser, et qu'elle prend en charge différents types de données telles que numériques, nominales, catégorielles, etc. Le but de l'utilisation d'un arbre de décision est de créer un modèle d'entraînement qui peut prédire la classe ou la valeur de la variable cible en fonction des règles de décision fondamentales apprises à partir de données précédentes (données d'entraînement).

Le résultat de la classification de l'arbre de décision est illustré dans la Figure IV.2. Nous avons obtenu neuf feuilles qui seront exploitées pour générer des règles SWRL qui seront utilisées dans le modèle d'ontologie. Figure IV.3 fournit un extrait de la sortie de l'arbre de décision.

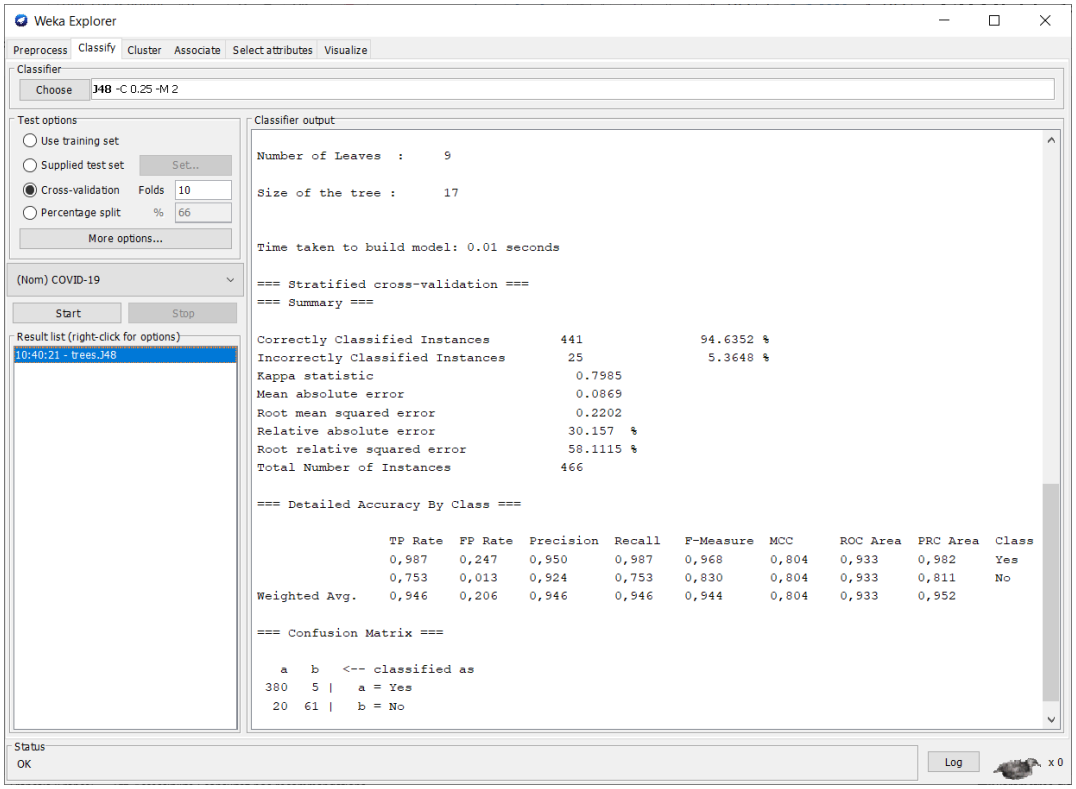

Figure IV.2 - Classification de l'arbre de décision.

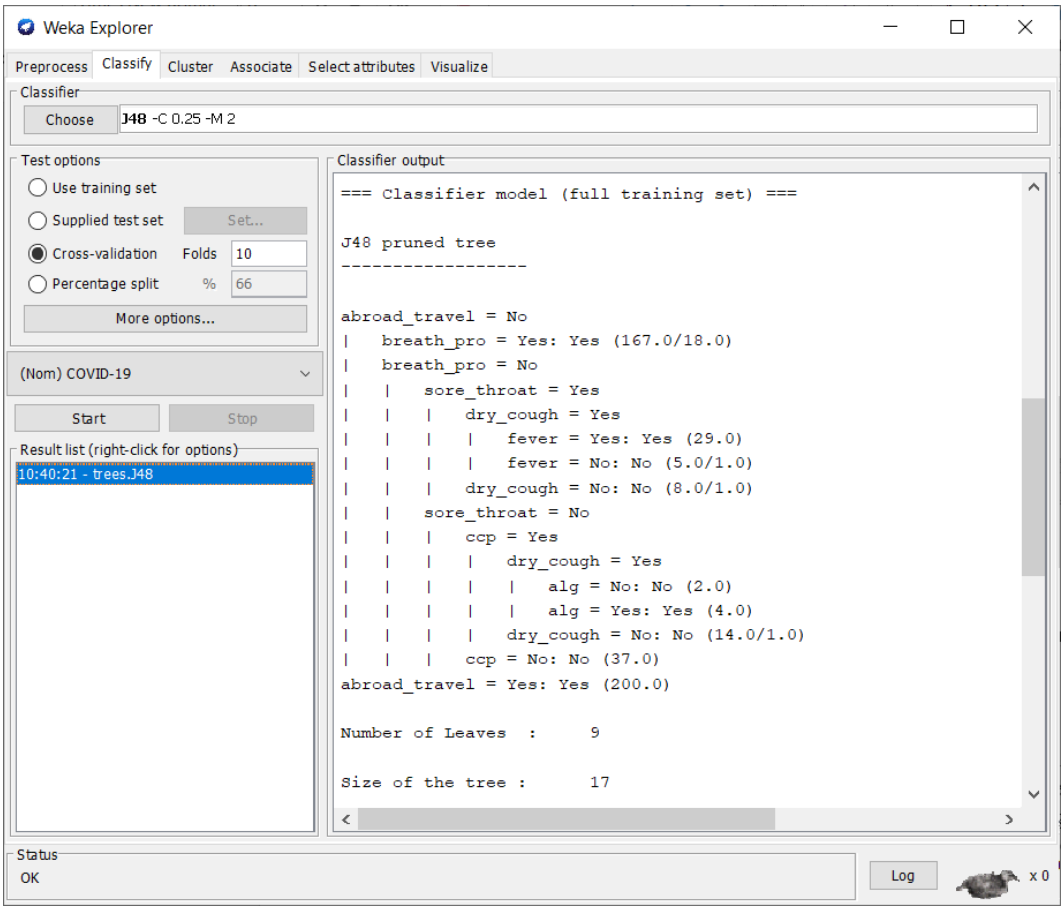

Figure IV.3 - Résultat de l'arbre de décision.

IV.3.3. Ingénierie ontologique et langage de règles du Web sémantique

Cette section présente les technologies utilisées pour créer l'ontologie, ainsi que l'approche utilisée pour construire le modèle d'ontologie à l'aide de règles extraites de l'algorithme d'arbre de décision.

L'ontologie a été construite à l'aide du logiciel Protégé, qui est une plate-forme open source qui fournit un ensemble d'outils à une communauté d'utilisateurs croissante pour construire des modèles de domaine et des applications basées sur les connaissances avec des ontologies. Nous avons créé l'ontologie et les classes principales de l'ontologie sont « Diagnostic » et « Patient ». Deux sous-classes (absence et présence) de la classe mère Diagnostic, et quatre sous-classes (TP, TN, FN, FP) de la classe mère « PatientEvaluationMetrics » qui est une sous-classe de « Patient ». La représentation graphique de l'ontologie est illustrée dans la Figure IV.4.

Les propriétés de données utilisées dans l'ontologie sont les mêmes attributs présentés dans la Table IV.1 qui sont utilisés pour construire des modèles d'algorithmes d'apprentissage automatique. La Figure IV.5 illustre les propriétés des données, et pour importer le même ensemble de données utilisé dans Weka, nous avons utilisé un plugin parmi les plugins du logiciel Protégé appelé Cellfie, qui permet d'alimenter les instances via un fichier CSV.

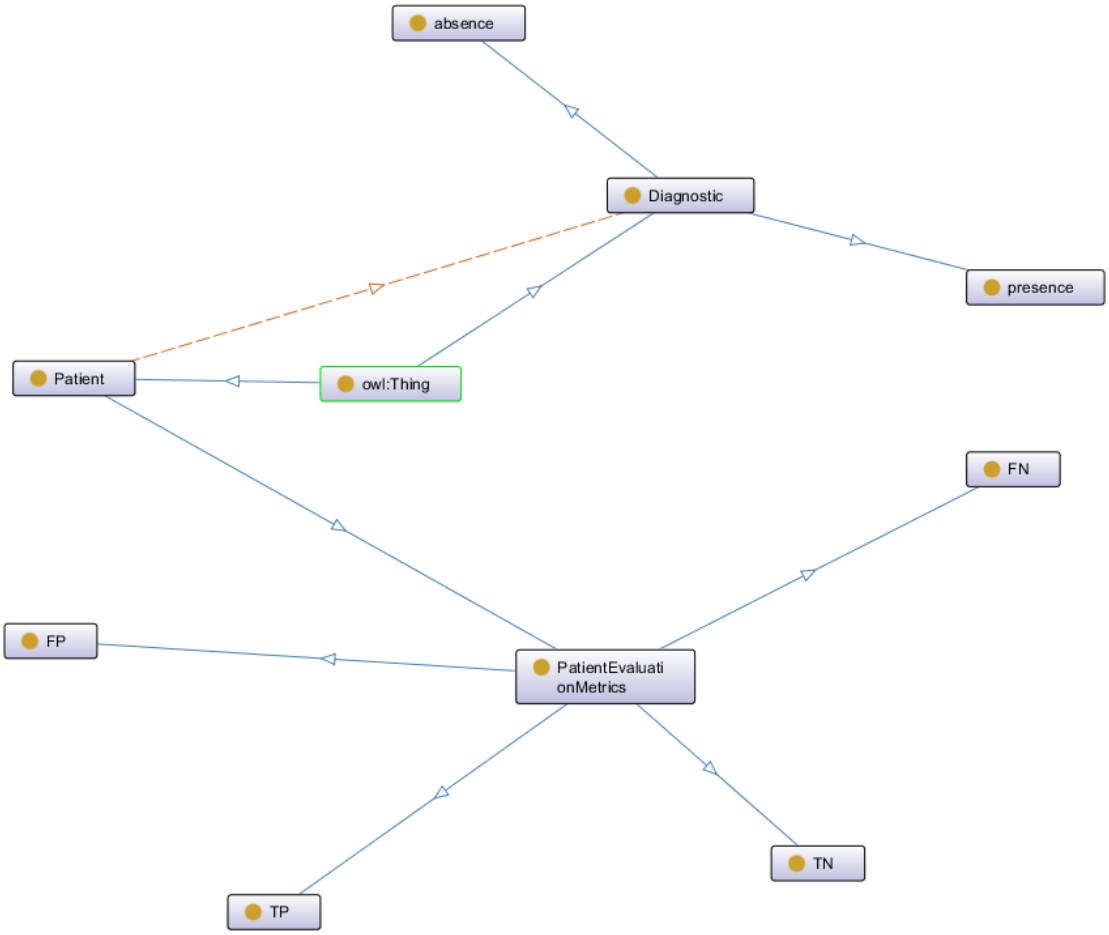

Figure IV.4 - Graphe ontologique.

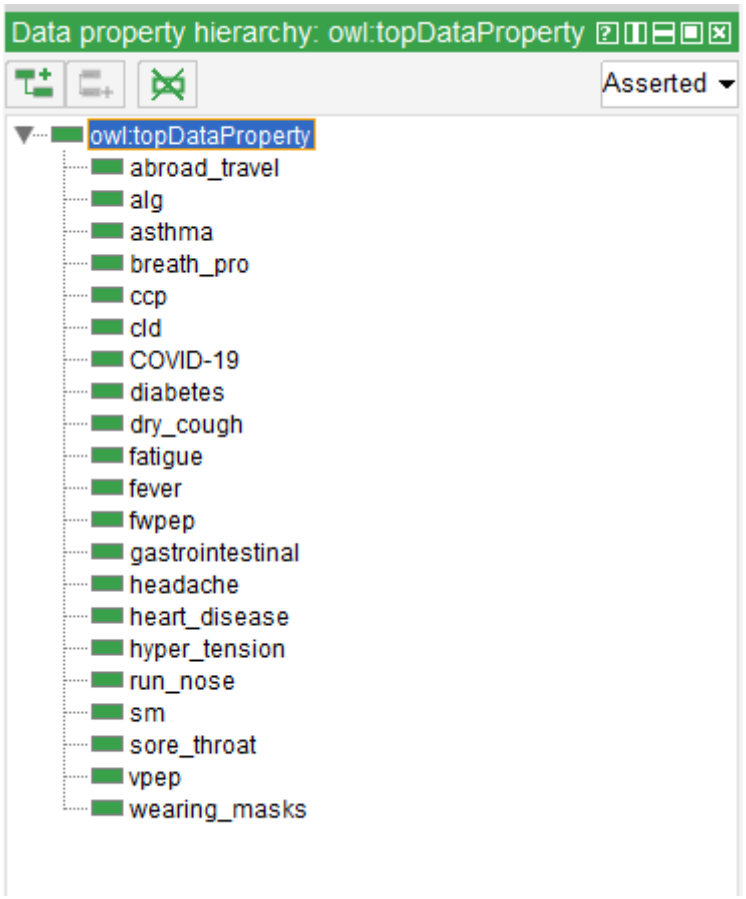

Figure IV.5 - Propriétés des données.

IV.3.3.1. Règles de langage du Web sémantique (SWRL) et raisonneur Pellet

Suite à la création de classes, de propriétés de données et d'instances dans l'ontologie. Nous devons établir les règles de raisonnement SWRL. Pour ce faire, nous avons utilisé le plugin SWRLTab, nous avons récupéré les règles créées à partir de l'algorithme de l'arbre de décision (Figure IV.3) et nous avons importé ces règles dans Protégé. Les règles collectées à partir de l'algorithme d'arbre de décision sont converties à l'aide du langage de programmation Java, chaque feuille de l'arbre étant extraite en tant que règle SWRL unique.

Exemple d'une feuille de l'algorithme de l'arbre de décision :

If breath_pro = No && sore_throat = Yes && dry_cough = Yes && fever = Yes THEN put the individual in presence

SWRL obtenu :

Patient(?pt) ^ breath_pro(?pt, ?Br) ^ swrlb:equal(?Br, 'No'^^xsd:string) ^ sore_throat(?pt, ?ST) ^ swrlb:equal(?ST, 'Yes'^^xsd: string) ^ dry_cough(?pt, ?DC) ^ swrlb:equal(?DC, 'Yes'^^xsd: string) ^ fever(?pt, ?F) ^ swrlb:equal(?F, 'Yes'^^xsd: string) → *presence*

Pour exécuter les règles SWRL et déduire de nouveaux axiomes d'ontologie, nous avons utilisé un autre plugin du logiciel Protégé nommé Pellet, qui inclut des capacités pour vérifier la cohérence de l'ontologie, traiter les règles SWRL, calculer la hiérarchie de classification, traiter avec OWL, expliquer les inférences et répondre aux requêtes SPARQL. Il implémente

les règles d'ontologie et de SWRL pour initier l'inférence, puis détermine si la présence ou l'absence de la maladie à coronavirus. Les résultats du classificateur d'ontologie sont rapportés dans la section suivante.

IV.3.4. Métriques d'évaluation

La zone ROC, la F-mesure, l'erreur quadratique moyenne (RMSE), le rappel, l'exactitude, l'erreur quadratique relative racine, la précision, la statistique kappa et d'autres mesures de performance sont utilisées pour évaluer les algorithmes d'apprentissage automatique. Nous avons utilisé deux modes de test (test fractionné et validation croisée K-fold) en utilisant plusieurs métriques, notamment le rappel, la F-mesure, la précision et la précision pour analyser nos résultats expérimentaux. De plus, les mêmes critères sont utilisés pour évaluer la validité de cette recherche comparative, y compris les classificateurs d'apprentissage automatique et le modèle ontologique. D'autres mesures, telles que l'erreur absolue moyenne (MAE), la MSE et la RMSE, sont disponibles mais sont le plus souvent utilisées dans les problèmes de régression. En conséquence, en raison des problèmes de classification imposés par l'ensemble de données et les techniques employées.

- **Exactitude (Accuracy) :**

L'exactitude est la mesure de toutes les instances correctement prédites sur le total des prédictions faites par le modèle, et chaque algorithme peut fonctionner différemment en ce qui concerne les instances correctement classées. L'exactitude calcule le rapport des instances correctement classées qui sont de vrais positifs (TP) et de vrais négatifs (TN) sur le nombre total de prédictions, y compris les TP et TN et les prédictions incorrectes, à savoir les faux positifs (FP) et les faux négatifs (FN). L'exactitude peut être calculée à l'aide de la formule suivante :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Précision (Precision) :**

De plus, les principales mesures de précision ont été incluses dans l'analyse comparative, telles que la précision, le rappel et la F-mesure. La précision mesure l'exactitude des prédictions de TP sur tous les positifs prédits en divisant le TP par la somme de TP et FP. Selon la description donnée, la précision signifie combien de personnes classées comme COVID-19-positives sont réellement COVID-19-positives, et elle peut être calculée à l'aide de cette formule :

$$Precision = \frac{TP}{TP + FP}$$

- **Rappel (Recall) :**

Le rappel mesure la précision de la prédiction de TP sur les instances positives réelles dans l'ensemble de données. Le rappel répond à la question, parmi toutes les instances positives au COVID-19, combien ont été correctement prédites par le modèle ? Le pourcentage de rappel peut être obtenu en divisant le TP par la somme de TP et FN.

$$Recall = \frac{TP}{TP + FN}$$

- **F-mesure (F-Measure)**

Étant donné que la précision et le rappel mesurent des choses différentes, la valeur de la « F-mesure » mesure l'harmonie, l'équilibre, des deux critères. Le score F-Measure diminuera si un critère est amélioré au détriment de l'autre, et il peut être calculé à l'aide de cette formule :

$$\text{F-Measure} = 2 * \frac{PREC * REC}{PREC + REC}$$

- **Instances classées correctement et incorrectement**

Ces valeurs ont également été prises en compte dans l'analyse comparative des algorithmes d'apprentissage automatique. Le résultat des instances correctement classées est la somme des prédictions TP et TN ; à l'inverse, le résultat des instances mal classées est la somme des prédictions FP et FN du modèle.

- **Kappa Statistic (K)**

La statistique kappa de Cohen calcule la fiabilité des résultats entre deux évaluateurs de la même chose ; c'est à quel point les évaluateurs sont d'accord par hasard. Un score de zéro signifie qu'il y a un accord aléatoire ou moins entre les deux évaluateurs, et le score peut être inférieur à zéro, alors qu'un score de 1 indique un accord complet. Il peut être calculé à l'aide de la formule suivante :

$$K = \frac{P_o - P_e}{1 - P_e}$$

Où P_o est la probabilité d'accord et P_e est la probabilité d'accord aléatoire entre les évaluateurs.

- **Erreur absolue moyenne (MAE) :**

Pour évaluer les performances du modèle, MAE est utilisé pour mesurer la quantité d'erreurs de classification ou d'erreurs dans la prédiction du modèle. MAE est la moyenne de toutes les erreurs absolues ; il détermine à quel point la valeur prédite est proche de la valeur réelle dans l'ensemble de données. MAE peut être obtenu par la formule suivante :

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

Où n représente le nombre total d'erreurs, Σ est le symbole de sommation, x_i est la valeur prédite, x est la valeur réelle et les barres verticales représentent la valeur absolue.

- **Erreur quadratique moyenne (MSE) :**

L'erreur quadratique moyenne est une autre façon de mesurer les performances des modèles de régression [283]. MSE prend la distance des points de données de la droite de régression et les met au carré. La quadrature est nécessaire car elle supprime le signe négatif de la valeur et donne plus de poids aux différences plus importantes. Plus l'erreur quadratique moyenne est

petite, plus vous vous rapprochez de la ligne de meilleur ajustement. MSE peut être calculé comme suit :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2$$

- **Erreur quadratique moyenne (RMSE) :**

L'erreur quadratique moyenne peut-être définie comme l'écart type des erreurs de prédiction. Les erreurs de prédiction, également appelées résidus, correspondent à la distance entre la ligne de meilleur ajustement et les points de données réels. La RMSE est donc une mesure de la concentration des points de données réels autour de la ligne de meilleur ajustement. Il s'agit du taux d'erreur donné par la racine carrée de MSE donnée comme suit.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2}$$

IV.4. Analyses des résultats et discussion

Dans cette section, les résultats de l'évaluation des différents classificateurs qui ont été utilisés dans cette étude sont présentés. Les statistiques et les résultats du modèle ontologique sont également présentés dans la Table IV.2, Table IV.3 et la Figure IV.6 illustrent les mesures de performance du modèle ontologique, la Figure V.6(a) représente une validation croisée de 10 fois et la Figure V.6(b) représente 70% mode fractionné.

Les résultats de cette étude fournissent une représentation visuelle des différentes mesures utilisées dans cette recherche, telles que la précision, la mesure F, le rappel et l'exactitude, comme le montrent les figures (Figure IV.7, Figure IV.8, Figure IV.9, Figure IV.10). La Table IV.4 montre également les résultats des différents classificateurs qui ont été utilisés dans cette recherche.

Table IV.2 - Validation croisée 10-fois pour le modèle ontologique.

Matrice de confusion		**Classe réelle**	
		positive	**négative**
Classe prédite	**positive**	TP : 389	FP : 10
	négative	FN : 2	TN : 65

Table IV.3 - Mode fractionné à 70% pour le modèle ontologique.

Matrice de confusion		**Classe réelle**	
		positive	**négative**
Classe prédite	**positive**	TP : 125	FP : 1
	négative	FN : 3	TN : 11

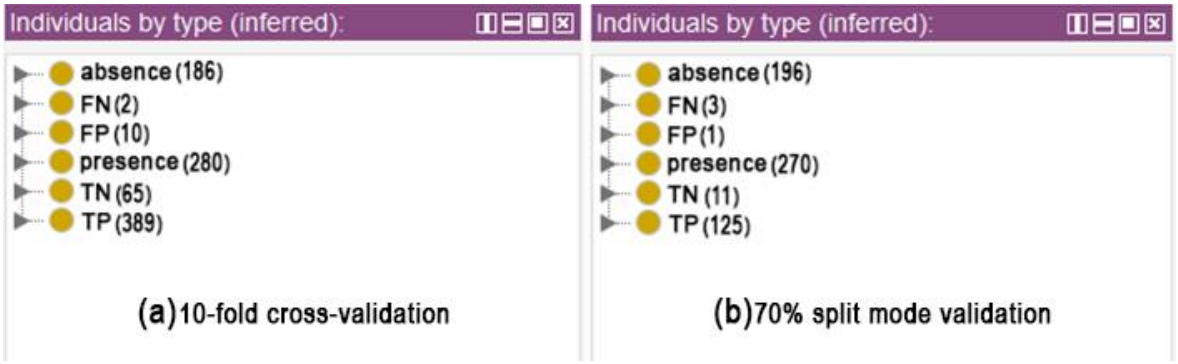

Figure IV.6 - Résultats des concepts inférés.

- **Exactitude :**

Selon la Figure IV.7 et la Table IV.4, concernant validation croisée de 10 fois le modèle ontologique a atteint la valeur maximale de 97,4 % et la machine à vecteurs de support avec un taux de 96,8 %, et 94,6 % pour l'arbre de décision et Naïve Bayes. Presque les mêmes résultats en utilisant le mode de test fractionné, nous avons obtenu 97,1 %, 96,4 % pour la machine à vecteur de support et 95,7 % pour la régression logistique et l'arbre de décision.

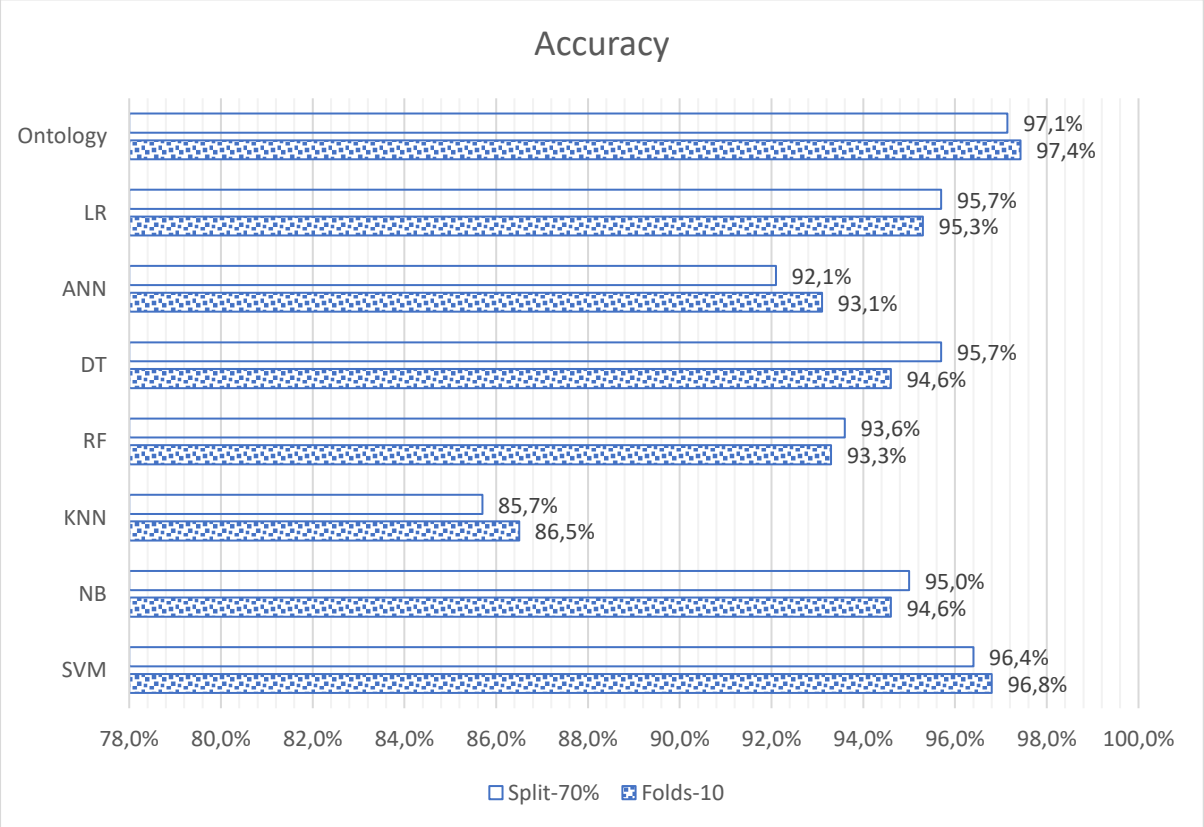

Figure IV.7 - Résultats de la comparaison de l'exactitude.

- **Précision :**

Le classificateur d'ontologie a la précision la plus élevée de 99,2 % en termes de mode de test fractionné, suivi de la machine à vecteur de support, de la régression logistique et de l'arbre

de décision. Concernant le mode de validation croisée 10 fois, la valeur de précision la plus élevée de 97,5 % est concerne le modèle d'ontologie. Plus de détails sont présentés dans la Table IV.4 et la Figure IV.8.

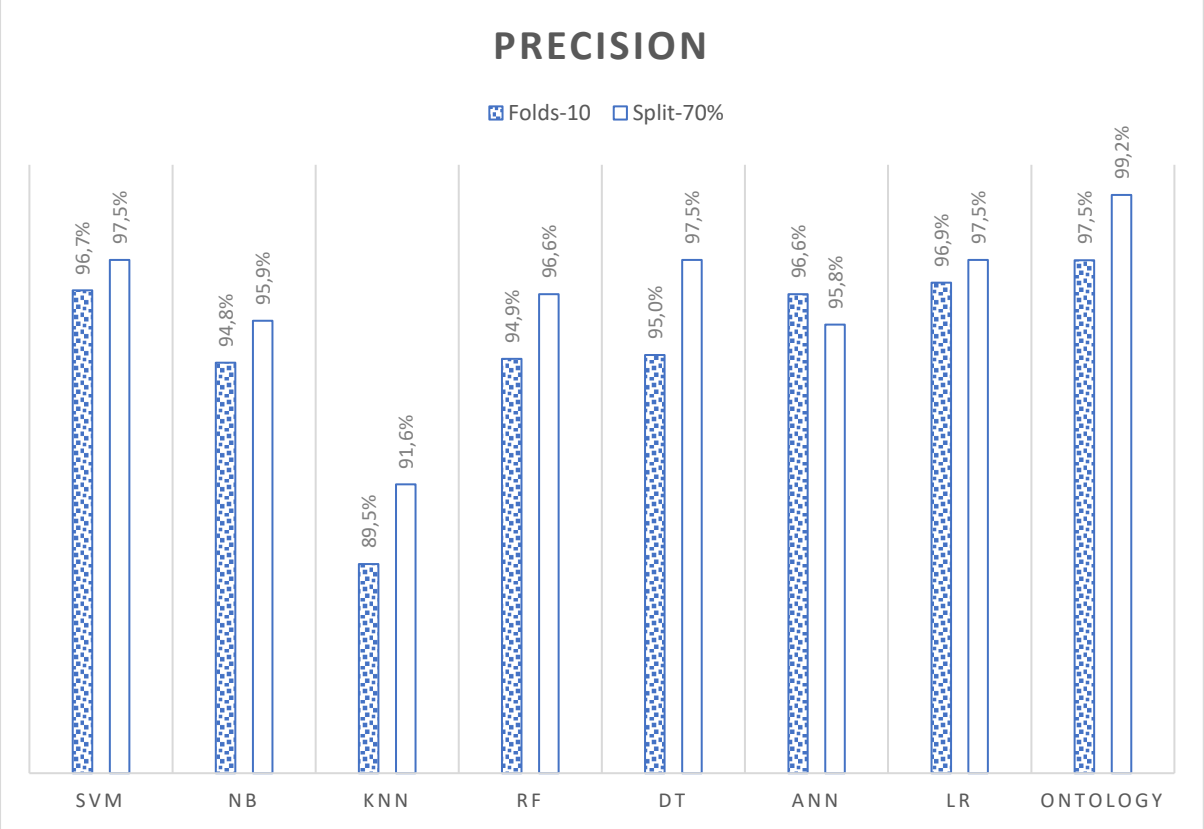

Figure IV.8 - Comparaison des résultats de précision.

- **Rappel :**

Selon la Figure IV.9 et la Table IV.4, le modèle ontologique et la machine à vecteur de support ont les valeurs de rappel les plus élevées de 99,5 % et 99,00 % pour Naïve Bayes qui concerne le mode de validation croisée 10 fois. En ce qui concerne le mode de test fractionné, la valeur de rappel la plus élevée de 98,3 % pour Naïve bayes et machine à vecteur de support.

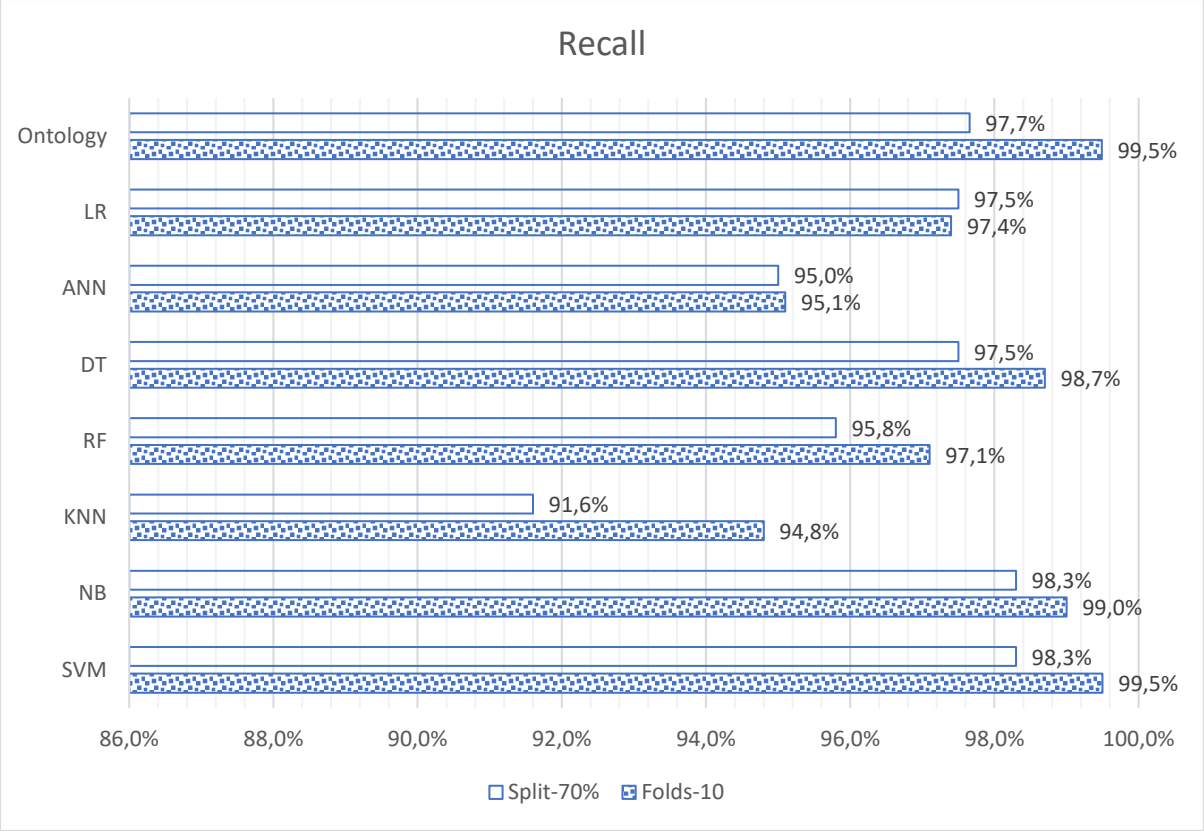

Figure IV.9 - Comparaison des résultats du rappel.

- **F-mesure :**

Selon la Figure IV.10 et la Table IV.4, le modèle d'ontologie avait la plus grande valeur de 98,5 % dans les deux modes de test, suivi de machine à vecteur de support en deuxième position et de la régression logistique en troisième position.

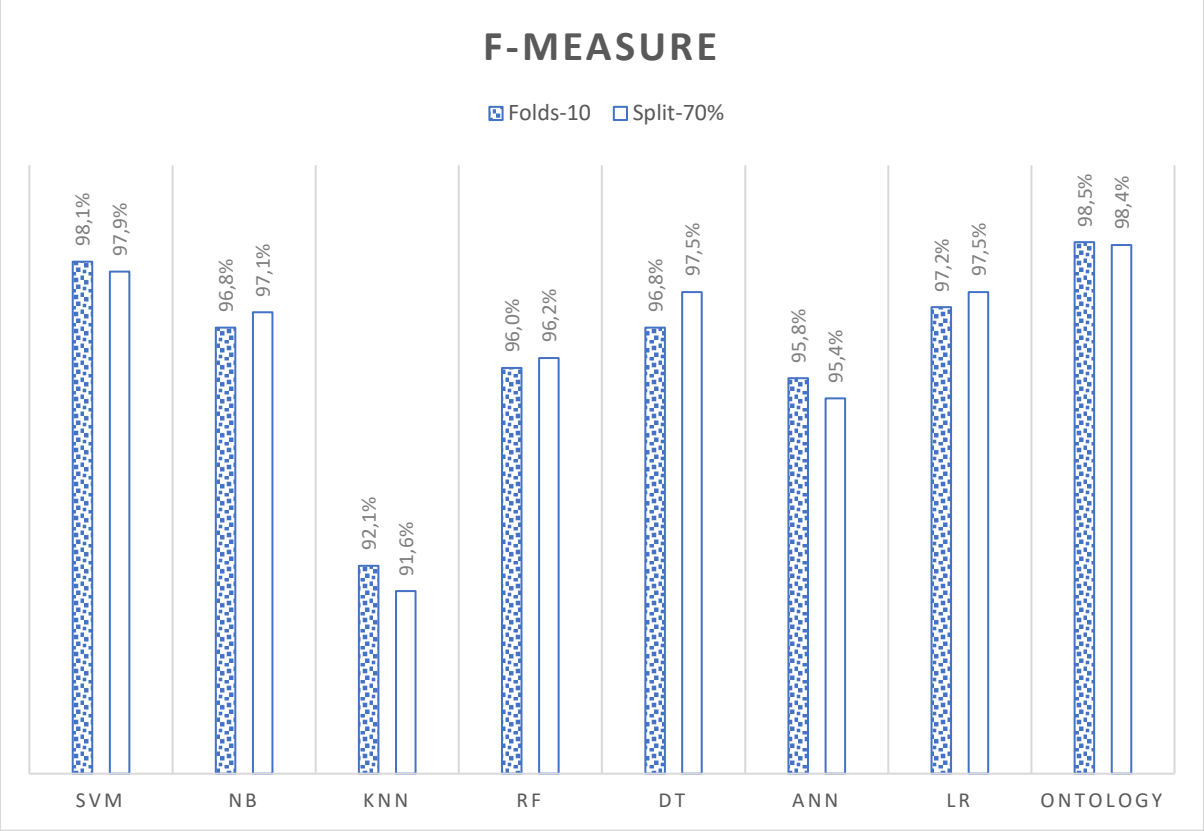

Figure IV.10 - Résultats de la comparaison F-mesure.

Les résultats expérimentaux révèlent que le modèle d'ontologie a la précision la plus élevée de 97,4 %, suivi de la machine à vecteur de support à 96,8 %, de la régression logistique à 95,3 % et de l'arbre de décision et de Naïve Bayes à 94,6 %. En termes de données indiquées ci-dessus, nous ne voyons aucune différence significative entre les modes de test 70%-Split et 10-Folds. Nous concluons que le modèle ontologique peut aider en étendant la portée du modèle d'apprentissage automatique. Ils peuvent comprendre n'importe quel type ou variation de données, et chaque donnée peut être affectée à un certain travail. La combinaison du modèle ontologique avec l'apprentissage automatique peut donner de bons résultats. Le modèle ontologique obtient des résultats comparables aux classificateurs d'apprentissage automatique. Les humains peuvent interpréter les résultats et les règles peuvent être modifiées ou ajoutées au besoin. De plus, il prend en charge les formats de données non structurés, semi-structurés et structurés, permettant une intégration plus transparente des données. Il peut comprendre tous les aspects du processus de modélisation des données, en commençant par les schémas au niveau le plus élémentaire. En conséquence, ils peuvent gérer les quantités massives de données utilisées comme entrées pour l'entraînement à l'apprentissage automatique ou les sorties comme résultats. De plus, l'ontologie correspond à l'objectif de toute organisation, qui peut être mathématique, logique ou sémantique. À notre connaissance, il s'agit de la première étude comparative du modèle ontologique et de l'apprentissage automatique dans laquelle nous avons intégré l'ontologie à l'apprentissage automatique, en particulier dans le domaine de la prédiction COVID-19.

Table IV.4 - Résultats du modèle ontologique et des classificateurs d'apprentissage automatique.

	Exactitude		**Précision**		**Rappel**		**F-mesure**	
	Folds-10	Split-70%	Folds-10	Split-70%	Folds-10	Split-70%	Folds-10	Split-70%
Support Vector Machine	0.968	0.964	0.967	0.975	0.995	0.983	0.981	0.979
Naïve Bayes	0.946	0.95	0.948	0.959	0.99	0.983	0.968	0.971
K-Nearest Neighbors	0.865	0.857	0.895	0.916	0.948	0.916	0.921	0.916
Random Forest	0.933	0.936	0.949	0.966	0.971	0.958	0.96	0.962
Decision Tree	0.946	0.957	0.95	0.975	0.987	0.975	0.968	0.975
Artificial Neural Network	0.931	0.921	0.966	0.958	0.951	0.95	0.958	0.954
Logistic Regression	0.953	0.957	0.969	0.975	0.974	0.975	0.972	0.975
Ontology Model	0.974	0.971	0.975	0.992	0.995	0.977	0.985	0.984

IV.5. Conclusion

La pandémie de COVID-19 a profondément marqué l'année 2020 et a fait réagir la communauté des chercheurs dans différents domaines. Cette étude visait à construire un modèle prédicteur de présence COVID-19 en appliquant sept algorithmes d'apprentissage automatique supervisés, notamment l'arbre de décision, algorithme de forêt aléatoire, méthode des k plus proches voisins, Naïve Bayes, machine à vecteur de support, régression logistique et réseau de neurones artificiels. Nous avons démontré comment les ontologies peuvent aider à prédire la présence de COVID-19 sur la base des symptômes, en intégrant l'ontologie et l'apprentissage automatique avec implémentation des règles de l'algorithme de l'arbre de décision dans le raisonneur d'ontologie. Une analyse comparative a été menée en évaluant les performances du modèle en validation croisée 10 fois et en fractionnement de test en pourcentage via le logiciel d'apprentissage automatique WEKA. Les résultats sont évalués à l'aide de mesures de performance générées à partir de la matrice de confusion, telles que la F-mesure, l'exactitude, la précision et le rappel. Selon les résultats, l'ontologie a dépassé tous les algorithmes d'apprentissage automatique avec une valeur de précision élevée de 97,4 %.

Cette étude peut servir de système d'aide à la décision pour les médecins, en utilisant le modèle développé comme aide pour détecter la présence de COVID-19 chez une personne en fonction des symptômes déclarés. De plus, les personnes qui présentent certains symptômes liés au COVID-19 peuvent l'utiliser pour déterminer la possibilité d'être positif ou négatif au test COVID-19. Cette étude peut encourager les individus à consulter immédiatement le médecin et favorise le diagnostic précoce de la maladie. De cette façon, il peut aider à prévenir la propagation de cette maladie contagieuse, en diminuant le danger pour les vies humaines. Le modèle développé dans cette étude peut être utilisé pour construire une application avec les avantages suivants :

- Les individus peuvent facilement vérifier la possibilité de contracter le COVID-19 en fonction des symptômes ;
- Cette étude peut être utilisée comme une évaluation préliminaire du patient pour les médecins praticiens ;
- Aider les entreprises à limiter les contacts physiques avec les clients potentiellement atteints de la COVID-19 ;
- Cette étude peut servir d'outil d'autogestion supplémentaire pour les installations de quarantaine afin de surveiller si la personne a développé des symptômes de la COVID-19 pendant son isolement ;
- La communauté et le gouvernement peuvent utiliser cette étude comme un outil pour réduire la propagation du virus grâce à la détection précoce de la COVID-19.

Cette étude sera continuellement améliorée dans le cours futur, nous prévoyons ensuite d'explorer la méthodologie de prédiction en utilisant l'ensemble de données mis à jour et d'utiliser les méthodes d'apprentissage automatique les plus précises et les plus appropriées pour la prévision. Les prévisions en temps réel seront l'un des principaux axes de nos travaux futurs.

CHAPITRE V - MODELE ONTOLOGIQUE BASE SUR L'APPRENTISSAGE AUTOMATIQUE POUR PREDIRE LE CANCER DU SEIN

V.1. Introduction

En 2020, 2,3 millions de femmes ont été identifiées avec un cancer du sein, avec 685 000 décès dans le monde. D'ici la fin de 2020, 7,8 millions de femmes auront reçu un diagnostic de cancer du sein au cours des cinq années précédentes, ce qui en fera le type de cancer le plus fréquent dans le monde. Le cancer du sein réclame plus de DALY (pour disability-adjusted life years) chez les femmes que tout autre type de cancer dans le monde. Le cancer du sein affecte les femmes de tous âges après la puberté dans tous les pays du monde, mais à un rythme croissant dans les dernières étapes de la vie.

Le cancer du sein est le premier cancer féminin dans le monde. Il est dû à la croissance anormale de certaines cellules du sein. Plusieurs techniques ont été introduites pour le diagnostic correct du cancer du sein. Le dépistage mammaire ou mammographie [284] est une technique de diagnostic du cancer du sein. Il est utilisé pour vérifier l'état du mamelon des femmes grâce aux rayons X. Généralement, il est presque impossible de détecter le cancer du sein au stade initial en raison de la petite taille de la cellule cancéreuse vue de l'extérieur. Il est possible de diagnostiquer le cancer à un stade précoce grâce à la mammographie, et ce test ne prend que quelques minutes.

Cette incidence de la maladie et les taux de mortalité varient selon la race et l'âge, cependant, elle est hautement guérissable lorsqu'elle est diagnostiquée tôt et avant qu'elle ne métastase [285]. Le diagnostic du cancer du sein est très difficile et fait l'objet d'une grande attention dans le monde entier en raison des conséquences associées à cette maladie car elle a des taux de morbidité et de mortalité élevés [286]. La prédiction de la catégorie de cancer au stade précoce est devenue un domaine essentiel de la recherche sur le cancer, car elle peut simplifier les exigences cliniques ultérieures des patients et déterminer les traitements efficaces [287]. Le diagnostic précoce du cancer du sein peut être un point déterminant entre la vie et la mort [288]. La technique traditionnelle pour diagnostiquer ce type de cancer consiste à utiliser l'imagerie par résonance magnétique (IRM) et l'examen microscopique du comportement de la tumeur pour déterminer le type de tumeur et si la tumeur est maligne ou bénigne. Une tumeur bénigne est un type de tumeur non invasive et elle provoque rarement des problèmes potentiellement mortels. En revanche, une tumeur maligne est un type invasif qui peut affecter les tissus environnants et métastaser dans des tissus distants du corps. Les approches modernes du diagnostic du cancer du sein utilisent l'apprentissage supervisé pour détecter les tumeurs avec une grande précision [289].

Le cancer du sein peut être classé comme bénin ou malin ; cependant, cette classification est déterminée par des tests de diagnostic. Certains critères à prendre en compte sont l'uniformité de la taille et de la forme des cellules, l'adhérence marginale, la taille des cellules épithéliales uniques, les noyaux nus, la chromatine fade et les nucléoles normaux. En observant ces critères, les médecins ou les scientifiques sont en mesure de poser un diagnostic en fonction des résultats des tests de diagnostic du patient.

Avec l'avancement des capacités et des technologies de pointe des domaines biomédicaux informatiques, de nombreux tests cliniques et informations sur les patients liés au cancer du sein ont été enregistrés. Pour contrôler l'augmentation rapide des cas de cancer du sein et minimiser les facteurs de risque, les chercheurs ont utilisé les dossiers cliniques historiques des patientes pour prédire le cancer du sein [290]–[293]. Une variété de modèles a été développés pour détecter le cancer à l'aide d'algorithmes d'apprentissage automatique tels que la régression logistique, l'arbre de décision, la forêt aléatoire, l'eXtreme Gradient Boosting (Xgboost), etc. [294].

Depuis quelques années, diverses techniques d'apprentissage automatique [295]–[298], d'apprentissage profond et d'informatique bio-inspirée sont utilisées dans plusieurs pronostics médicaux. Bien qu'un certain nombre de modalités aient été démontrées, aucune des modalités n'est en mesure de fournir un résultat correct et cohérent.

En plus d'identifier le meilleur modèle de classificateur qui introduit une plus grande précision de classification pour l'ensemble de données prédéfini utilisé dans cette étude, le processus de classification des données est mis en œuvre en appliquant des opérations de prétraitement et en extrayant des caractéristiques aux enregistrements de données spécifiés à partir de l'ensemble de données à l'aide de WEKA. Le WEKA (Waikato Environment for Knowledge Analysis) est un logiciel open-source qui contient un ensemble d'algorithmes pour les tâches d'exploration de données [299]. Ces algorithmes peuvent être appliqués à un ensemble de données soit directement via l'interface WEKA, soit via du code Java. Ensuite, les différents classificateurs sont implémentés avec différentes variables en utilisant plusieurs algorithmes et plusieurs options pour calculer le meilleur rapport de l’exactitude.

Dans cette contribution, nous avons l'intention d’intégrer un modèle ontologique pour prédire le cancer de sein en comparaison avec sept approches de classification importantes en utilisant des critères soigneusement choisis obtenus à partir de la matrice de confusion, tels que la F-mesure, l'exactitude, la précision et le rappel. Le reste de ce chapitre est organisée comme suit. La section suivante décrit l'état actuel de l'état de l'art dans ce domaine, puis illustre les méthodes et les matériaux utilisés pour l'étude. Le concept théorique de chaque technique d'apprentissage automatique et l’ontologie est illustré dans la section suivante. Ensuite, les paramètres de mesure des performances sont décrits. La configuration expérimentale et l'analyse des résultats sont étudiées avant la section finale. La dernière section présente la conclusion du chapitre.

V.2. Travaux connexes

Avec l'évolution de la recherche médicale, de nombreux nouveaux systèmes ont été développés pour la détection du cancer du sein. La recherche associée à ce domaine est brièvement décrite ci-dessous.

Dans cette étude [46], les auteurs se sont concentrés sur l'amélioration de la valeur de l’exactitude à l'aide d'un algorithme de sélection de caractéristiques appelé particle swarm optimization (PSO) ainsi que des algorithmes d'apprentissage automatique K-Nearest Neighbors (KNN), Naive Bayes (NB) et reduced error pruning (REP) tree. Leur perspective de travail porte sur le problème du cancer du sein chez les femmes saoudiennes et, selon leur rapport, c'est l'un des problèmes majeurs en Arabie saoudite. Leurs rapports suggèrent que les

femmes ayant une tranche d'âge supérieure à 46 ans sont les principales victimes de cette maladie malveillante. Tenant compte de ce sentiment, ils ont mis en œuvre cinq techniques d'analyse de données basées sur les phases sur l'ensemble de données WBCD. Ils ont rapporté une analyse comparative entre la classification sans méthode de sélection des caractéristiques et la classification avec une méthode de sélection des caractéristiques. Ils ont acquis une précision de 70 %, 76,3 % et 66,3 % pour NB, RepTree et K-NN, respectivement. Ils ont utilisé l'outil Weka à des fins d'analyse de données. Avec la mise en œuvre de PSO, ils ont trouvé quatre fonctionnalités qui conviennent le mieux à cette tâche de classification. Pour NB, RepTree et K-NN avec PSO, ils ont obtenu respectivement des valeurs de précision de 81,3 %, 80 % et 75 %.

[300] Les auteurs ont proposé une technique d'arbre de décision modifiée en tant qu'arbre de décision à poids amélioré et l'ont implémentée sur WBCD et un autre ensemble de données sur le cancer du sein extrait du référentiel UCI. À l'aide du test Chisquare, ils ont classé chaque caractéristique et ils ont conservé les caractéristiques pertinentes pour cette tâche de classification. Pour l'ensemble de données WBCD, leur technique proposée a acquis une précision d'environ 99 %, tandis que pour l'ensemble de données sur le cancer du sein, elle a acquis une précision d'environ 85 à 90 %.

Les auteurs de cette recherche [301], ont principalement présenté des revues complètes sur les techniques Support Vector Machine, K-Nearest Neighbors, Artificial Neural Network et Decision Tree dans l'application de la prédiction du cancer du sein sur l'ensemble de données de référence du diagnostic du cancer du sein du Wisconsin (WBCD). Selon les auteurs, l'approche des réseaux de croyance profonde (DBN) avec l'architecture ANN (DBN-ANN) a donné le résultat le plus précis. Cette architecture a obtenu une précision de 99,68 %, alors que pour la méthode SVM, l'algorithme de clustering en deux étapes associées à la technique SVM a atteint une précision de classification de 99,10 %. Ils ont également examiné la technique d'ensemble où SVM, Naive Bayes et J48 ont été mis en œuvre à l'aide de la technique de vote. La méthode d'ensemble a acquis une exactitude de 97,13 %.

Les auteurs dans [302] ont mis l'accent sur les techniques Naive Bayes sur la prédiction du cancer du sein et décrit une étude comparative sur Tree Augmented Naive Bayes (TAN), Boosted Augmented Naive Bayes (BAN) et Bayes Belief Network (BBN). Ils ont utilisé SAS-EM (Statistical Analytical Software Enterprise Miner) pour la mise en œuvre des modèles. Le même ensemble de données WBCD populaire est utilisé dans leur travail. Selon leurs conclusions, à l'aide de l'amplification du gradient, une précision de 91,7 %, 91,7 % et 94,11 % a été obtenue pour BBN, BAN et TAN, respectivement. Par conséquent, leurs recherches suggèrent que TAN est le meilleur classificateur parmi les techniques Naive Bayes pour cet ensemble de données.

Dans cette étude [303], les auteurs ont introduit une méthode de prédiction du cancer du sein en utilisant les variantes de l'arbre de décision. Les modalités utilisées dans cette technique sont l'arbre de décision unique (SDT), l'arbre de décision boosté (BDT) et la forêt d'arbres de décision (DTF). La décision est prise en formant l'ensemble de données et après ce test. Les résultats ont présenté que l'exactitude obtenue par SDT et BDT est de 97,07 % et 98,83 %, respectivement, dans la phase d'entraînement, ce qui clarifie que BDT a mieux performé que SDT. La forêt d'arbres de décision a obtenu une exactitude de 97,51% alors que SDT 95,75% dans la phase de test. L'ensemble de données a été formé par un mode de validation croisée

décuplé. Dans [304], les auteurs ont mis en évidence une procédure de détection du cancer du sein. Les expériences qui ont été faites pour détecter la maladie sont discutées ici en utilisant le réseau neuronal à ondelettes linéaire local (LLWNN) et les moindres carrés récursifs (RLS) pour améliorer les performances du système. Le LLWNN-RLS fournit les valeurs maximales du taux de classification correcte (CCR) moyen de 0,897 et 0,972 pour 2 et 3 prédicteurs, respectivement, avec quelques temps de calcul. Il fournit également la valeur la plus basse de la longueur de description minimale (MDL) et de l'erreur quadratique moyenne de classification (ASCE) avec beaucoup moins de temps.

Le papier [305], a proposé un système hybride pour la détection du cancer du sein en utilisant KPSO et RLS pour RBFNN. Les centres, ainsi que les variances de RBFNN, sont ajustés à l'aide de K-particle swarm optimization et ajustés à l'aide de back-propagation. L'exactiude de classification obtenue par RBFNNKPSO et le filtre de Kalman étendu RBFNN est de 97,85 % et 96,42 %, respectivement, tandis que le temps de couverture est de 8,38s et 4,27s, respectivement. Les auteurs de [306], ont développé un modèle mathématique pour la prédiction du cancer du sein basé sur la régression symbolique de la programmation génétique multigénique. La technique du décuplement est utilisée pour éviter le surajustement. Une étude comparative est également illustrée. Les critères d'arrêt du modèle ont été générés mais le niveau de génération n'a pas atteint zéro. L'exactitude la plus élevée obtenue par le modèle est de 99,28 % avec une précision de 99,26 %. Une variante de la machine à vecteurs de support [307] est introduite pour le diagnostic du cancer du sein. Ici, six types de SVM sont expliqués et utilisés pour l'évaluation des performances. Les résultats SVM standard sont comparés à d'autres types de SVM. La quadruple validation croisée est utilisée pour l'entraînement et les tests. L'exactitude, la spécificité et la sensibilité les plus élevées obtenues par St-SVM sont de 97,71 %, 98,9 % et 97,08 %, respectivement, dans la phase d'entraînement. L'exactitude, la sensibilité et la spécificité les plus élevées obtenues par NSVM, LPSVM, SSVM et LPSVM sont de 96,5517 % 98,2456 %, 96,5517 % et 97,1429 % individuellement lors de la phase de test.

Les auteurs dans [308], ont présenté une méthode efficace pour la détection du cancer du sein en catégorisant les attributs de l'ensemble de données du cancer du sein en utilisant la technique de programmation logique inductive. Une étude de comparaison avec un classificateur propositionnel est également réalisée. Les statistiques Kappa, la mesure F, l'aire sous la courbe ROC, le taux de vrais positifs, etc. sont calculés comme une mesure de performance. Le système a été simulé sur deux plates-formes nommées Aleph et WEKA. Les auteurs de cette étude [309], ont évalué des variantes d'algorithmes d'arbre de décision pour le diagnostic du cancer du sein. Le système a utilisé les algorithmes d'arbre de décision les plus courants nommés CART et C4.5 qui sont simulés dans la plate-forme WEKA à l'aide de Matlab et Python. Le CART implémenté en Python a atteint la précision la plus élevée de 97,4 % et la sensibilité la plus élevée de 98,9 % est obtenue dans le CART qui est implémenté dans Matlab, et une spécificité de 95,3 % est acquise par CART et C4.5, respectivement, qui sont simulés dans WEKA.

Les comparaisons effectuées dans cette étude [310], étaient basées sur les performances de quatre algorithmes d'apprentissage automatique différents : Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (C4.5) et k-Nearest Neighbors (k- NN) et ont été menées sur les ensembles de données Wisconsin Breast Cancer (WBC). L'objectif de l'étude

et des expériences menées était de déterminer l'efficacité de chaque algorithme en termes de précision, d'exactitude, de spécificité et de sensibilité. Les résultats ont montré que SVM a obtenu une exactitude de 97,13 % et a surpassé l'algorithme Naïve Bayes, C4.5 et k-Nearest Neighbors (k-NN) qui a obtenu une variance d'exactitude comprise entre 95,12 % et 95,28 %.

Dans [29], l'algorithme génétique (AG) a été utilisé parallèlement à différentes techniques d'exploration de données pour WBC. GA a été utilisé pour extraire des caractéristiques significatives et informatives afin de réduire la complexité de calcul et d'améliorer la vitesse de traitement de l'exploration de données. Les techniques d'exploration de données utilisées dans cette étude étaient les suivantes : arbres de décision (DT), réseau bayésien (BN), régression logistique (LR), forêt aléatoire (RF), SVM, forêt de rotation, réseaux à fonction de base radiale (RBFN) et Perceptron multicouche (MLP). Deux ensembles de données médicales WBC (WBC et Wisconsin Diagnostic Breast Cancer (WDBC)) ont été utilisés pour tester les performances des modèles algorithmiques. L'exactitude la plus élevée de 99,48 % a été obtenue par la sélection des fonctionnalités Random Forest et GA.

L'étude menée par [311] visait à diagnostiquer le cancer du sein en utilisant trois techniques différentes, à savoir : SVM, DT et Artificial Neural Network (ANN). L'étude a été appliquée sur l'ensemble de données WDBC de l'UCI. La sélection des caractéristiques a été appliquée pour augmenter l'efficacité des méthodes. La méthode d'ensemble a donné les meilleurs résultats parmi les méthodes utilisées. Il a donné une exactitude de 98,77 %, une sensibilité de 98,05 % et une spécificité de 100 %.

Dans [312], les auteurs ont utilisé trois méthodes d'exploration de données bien connues, à savoir, Naïve Bayes (NB), J48 et RBF Network pour développer des modèles de prédiction de la survie au cancer du sein. Les données, qui contiennent 683 instances, ont été acquises auprès de l'UCI. Pour développer les modèles de prédiction, la sélection, le prétraitement et la transformation des données ont été appliqués. Les résultats obtenus à partir de l'expérience ont montré que le Naïve Bayes a donné les meilleurs résultats avec une exactitude de classification de 97,36 %, RBF Network a entraîné une exactitude de classification de 96,77 % et le J48 a entraîné une exactitude de classification de 93,41 %.

Les travaux de [313], ont utilisé douze techniques d'apprentissage automatique différentes pour le diagnostic du cancer du sein. Les techniques qui ont été utilisées sont à savoir ; NB, Decision Table, Ada Boost M1, J48, J-Rip, Régression logistique, Lazy IBK, Lazy K-star, Multiclass Classifier, Multilayer–Perceptron, RF et RT. Le jeu de données WBCD a été utilisé pour entraîner le modèle. La plupart des méthodes appliquées ont obtenu un score supérieur à 94 %. Seul NB a sous-performé, par rapport aux autres modèles, avec une exactitude de 73,21 %. Les algorithmes de classification RT et Lazy ont surpassé les autres avec une exactitude proche de 99 %.

Dans [314], les chercheurs ont utilisé huit techniques d'exploration de données différentes pour la prédiction du cancer du sein. Le jeu de données utilisé pour l'expérience était WPBC. Les expérimentations ont été faites sur quatre algorithmes de classification : SVM, DT C5.0, NB et k-NN et sur quatre algorithmes de clustering : EM, K means, PAM et Fuzzy c-means. Les expériences ont été réalisées à l'aide de l'outil de programmation R. Les résultats ont montré que les algorithmes de classification ont de meilleures performances que le clustering

où SVM et DT (C5.0) avaient la meilleure exactitude de 81 % et Fuzzy c-means a entraîné l'exactitude la plus faible de 37 %, parmi les algorithmes testés.

L'étude menée par [38] a proposé des méthodes d'ensembles nichés pour distinguer entre les tumeurs bénignes et les malins. Chaque méthode d'ensemble contient des "classificateurs", ainsi que des "métaclassificateurs" qui peuvent avoir plus de deux algorithmes de classification. Des métaclassificateurs ont été développés dans l'ensemble imbriqué à deux couches. Le jeu de données utilisé pour les expériences était WBDC. La méthode proposée a été comparée aux classificateurs simples conventionnels tels que BN et NB. Les résultats ont indiqué que la méthode d'ensemble imbriqué à deux couches surpasse les classificateurs simples.

Pour analyser les données sur le cancer du sein, [315] les auteurs ont utilisé quatre algorithmes de classification d'arbre de décision différents, à savoir les arbres de classification et de régression (CART), J48, Best First Tree (BF Tree) et DT (AD Tree). L'expérience a utilisé l'outil WEKA et les résultats ont démontré que le classificateur J48 atteignait l'exactitude la plus élevée de 99 %, tandis que les algorithmes CART obtenaient une exactitude de 96 % ; L'algorithme AD Tree a donné 97% et l'algorithme BF Tree a donné 98%.

Dans le papier [316], les auteurs ont utilisé neuf algorithmes de classification d'apprentissage automatique pour l'apprentissage supervisé (SL) et semi-supervisé (SSL) : 1) Régression logistique ; 2) Bayes naïf gaussien ; 3) Machine vectorielle à support linéaire ; 4) machine vectorielle de soutien RBF ; 5) Arbre de décision ; 6) Forêt aléatoire ; 7) Xgboost ; 8) Amplification de gradient ; 9) KNN. L'ensemble de données Wisconsin Diagnosis Cancer a été utilisé pour entraîner et tester ces modèles. Pour assurer la robustesse du modèle, nous avons appliqué une validation croisée K-fold et des hyperparamètres optimisés. Ils ont évalué et comparé les modèles en utilisant l'exactitude, la précision, le rappel, le score F1 et la courbe ROC. Les résultats de tous les modèles sont inspirants en utilisant à la fois SL et SSL. Le SSL a une grande précision (90 % à 98 %) avec seulement la moitié des données d'entraînement. Le modèle KNN pour le SL et la régression logistique pour le SSL ont atteint l'exactitude la plus élevée de 98 %.

L'étude [317] a été menée à l'aide de l'ensemble de données BCSC qui comprenait 280 660 résultats de mammographie de dépistage et les profils démographiques des patientes atteintes d'un cancer du sein qui sont des femmes âgées de 35 ans et plus. Les auteurs tentent d'appliquer trois techniques d'équilibrage de classe différentes, à savoir le suréchantillonnage (Synthetic Minority Oversampling Technique (SMOTE)), le sous-échantillonnage (SpreadSubsample) et une méthode hybride (SMOTE et SpreadSubsample) sur l'ensemble de données du Breast Cancer Surveillance Consortium (BCSC) avant de construire le système supervisé. Les algorithmes d'apprentissage utilisés dans cette étude incluent Naïve Bayes, Bayesian Network, Random Forest et Decision Tree (C4.5). Les résultats ont montré que le réseau bayésien généré à partir des données BCSC équilibrées en classe en utilisant la méthode hybride avait une meilleure performance globale en termes de ROC (0,937), de sensibilité (78,1 %), et de taux de faux positifs (0 %) ou de spécificité (100 %). Cette étude prouve que le modèle de réseau bayésien peut servir de meilleur système d'aide à la décision pour les médecins et de moyen de diagnostic et de traitement précoces pour les patientes en prédisant l'apparition du cancer du sein en fonction des facteurs de risque.

Dans cet article [318], un ensemble de données expérimentales universitaires sur le cancer du sein est utilisé pour effectuer une expérience pratique d'exploration de données à l'aide de l'outil Waikato Environment for Knowledge Analysis (WEKA). L'application Java WEKA représente une ressource riche pour réaliser des métriques de performance lors de l'exécution d'expériences. Le prétraitement et l'extraction de caractéristiques sont utilisés pour optimiser les données. Le processus de classification utilisé dans cette étude a été résumé à travers treize expériences. De plus, 10 expériences utilisant divers algorithmes de classification différents ont été menées. Les algorithmes introduits étaient : Naïve Bayes, Logistic Regression, Lazy IBK (Instance-Bases learning with parameter K), Lazy Kstar, Lazy Locally Weighted Learner, Rules ZeroR, Decision Stump, Decision Trees J48, Random Forest et Random Trees. Le processus de production d'un modèle prédictif a été automatisé grâce à l'utilisation de la précision de la classification. En outre, plusieurs expériences sur la classification du cancer du sein diagnostique du Wisconsin et du cancer du sein du Wisconsin ont été menées pour comparer les taux de réussite des différentes méthodes. Les résultats concluent que le classificateur Lazy IBK k-NN peut atteindre une exactitude de 98 % par rapports aux autres classificateurs. Les principaux avantages de l'étude étaient la compacité de l'utilisation de 13 modèles d'exploration de données différents et de 10 mesures de performance différentes, et le traçage des chiffres des erreurs de classification.

Dans cette étude [319], l'analyse discriminante linéaire et la machine à vecteurs de support sont comparées, selon le résultat, la machine à vecteurs de support et l'analyse discriminante linéaire ont de bonnes performances basées sur l'exactitude, la sensibilité, la spécificité et le score F1. En comparant les deux méthodes basées sur le nombre de résultats, les auteurs ont conclu que la machine à vecteurs de support mieux que l'analyse discriminante linéaire, avec un taux d'exactitude élevé de 98,77%. Les machines à vecteurs de support ont été largement utilisées par les chercheurs notamment sur la classification du cancer du sein car elles ont de bonnes performances. La machine à vecteurs de support est suggérée pour aider le médecin à prédire et à classer une maladie ou un ensemble de données similaires.

Cet article [320] présente l'approche bayésienne complète pour évaluer la distribution prédictive de toutes les classes à l'aide de trois classificateurs ; naïve bayes (NB), réseaux bayésiens (BN) et Tree Augmented Naïve Bayes (TAN) avec trois jeux de données ; Cancer du sein, cancer du sein wisconsin et ensemble de données sur les tissus mammaires. Ensuite, les exactitudes de prédiction des approches bayésiennes sont également comparées à trois algorithmes d'apprentissage automatique standard de la littérature ; K-plus proche voisin (K-NN), machine à vecteurs de support (SVM) et arbre de décision (DT). Les résultats ont montré que la meilleure performance était l'algorithme des réseaux bayésiens (BN) avec une précision de 97,281 %.

Le travail [321] implique la segmentation du noyau et la classification des caractéristiques du noyau prédit pour l'obtention de la meilleure précision. La carte des caractéristiques extraites à l'aide du réseau pyramidal des caractéristiques a été modifiée à l'aide de la segmentation des objets par convolution d'emplacements (SOLO) avec optimisation de la sauterelle pour la classification multiclasse. Une technique de multi-classification du cancer du sein basée sur un algorithme d'apprentissage en profondeur suggéré a été examinée pour atteindre une précision de 99,2 % à l'aide d'une énorme base de données de l'ICIAR 2018, démontrant l'efficacité de la méthode pour la multi-classification du cancer du sein dans un cadre médical. L'exactitude de segmentation obtenue est de 88,46 %.

L'étude [322] vise à développer un système de détection assistée par ordinateur utilisant l'apprentissage automatique à des fins de classification. Dans ce travail, 80 mammographies numériques de seins normaux, 40 de cas bénins et 40 de cas malins ont été choisis à partir du jeu de données mini MIAS. Ces images ont été débruitées à l'aide d'un filtre médian après avoir été segmentées pour obtenir une région d'intérêt (ROI) et améliorées à l'aide de l'égalisation d'histogramme. Ce travail fait une comparaison entre les performances du réseau de neurones artificiels (ANN), de la machine à vecteurs de support (SVM), des caractéristiques réduites de SVM et de l'hybride SVM-ANN pour le processus de classification à l'aide des caractéristiques statistiques et de la matrice de cooccurrence de niveaux de gris (GLCM) extraites des images améliorées. On constate que l'hybride SVM-ANN donne la meilleure précision de 99,4% et 100% pour différencier respectivement les cas normaux des cas anormaux (des cas bénins et des cas malins). Ce modèle hybride SVM-ANN a été déployé dans le développement du système CAO qui a montré une exactitude relativement bonne de 98 %.

Dans ce travail [323], l'ensemble de données sur le cancer du sein du Wisconsin a été utilisé, qui a été collecté à partir du référentiel UCI. L'objectif de cette étude est d'analyser l'ensemble de données et d'évaluer les performances de divers algorithmes d'apprentissage automatique pour prédire le cancer du sein. Ici, les classificateurs Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Decision Tree, Naive Bayes et Random Forest ont été mis en œuvre pour classer les tumeurs en deux classes bénignes et malignes. La précision de chaque algorithme est calculée et comparée pour trouver celui qui convient le mieux. Sur la base de l'analyse, Random Forest et Support Vector Machine surpassent les autres classificateurs avec une précision de 96,5 %. Ces classificateurs peuvent être utilisés pour construire un système de diagnostic automatique pour le diagnostic préliminaire du cancer du sein.

Dans cet article [324], l'étude de l'identification du cancer du sein est réalisée à l'aide de diverses techniques d'apprentissage automatique. Les résultats précis peuvent être utiles pour prédire un cancer malin ou bénin. La technique d'apprentissage automatique et la vision par ordinateur sont appliquées pour extraire les caractéristiques et développer le modèle d'optimisation en faisant des hyperparamètres de certaines valeurs. L'analyse est effectuée à l'aide de la machine à vecteurs de support et diverses mesures de qualité sont calculées telles que la précision, le score F1 et l'exactitude. Les résultats obtenus sont importants.

L'objectif de l'étude [325] est d'évaluer l'exactitude de prédiction des algorithmes de classification en termes d'efficience et d'efficacité. Les auteurs ont fourni une analyse détaillée des algorithmes de classification tels que Support Vector Machine, J48, Naïve Bayes et Random Forest en termes de précision de prédiction en appliquant une technique de validation croisée de 10 fois sur l'ensemble de données Wisconsin Diagnostic Breast Cancer à l'aide de l'outil open source WEKA. Le résultat de cette étude indique que Support Vector Machine a atteint l'exactitude de prédiction la plus élevée de 97,89 % avec un faible taux d'erreur de 0,14 %.

Dans cette recherche [326], deux algorithmes d'apprentissage automatique, à savoir le classificateur d'arbre de décision et la régression logistique, sont mis en œuvre pour la prédiction du cancer du sein, et ont comparé leurs exactitudes pour trouver lequel des deux sera le mieux adapté à la prédiction. Les résultats montrent que le classificateur d'arbre de

décision est l'algorithme le mieux adapté pour la prédiction, car son utilisation avait une exactitude de prédiction précise sur « Breast Cancer Wisconsin (Diagnostic) Data Set ».

L'étude [327] démontre l'utilisation d'arbres de décision pour représenter le diagnostic réel du cancer du sein pour le traitement local et systémique, ainsi que des stratégies supplémentaires qui peuvent être employées. L'étude évalue les performances de l'algorithme de l'arbre de décision en termes de précision de la classification, à l'aide de plusieurs mesures de précision telles que la mesure F, la zone ROC, la précision, le rappel, le taux TP et le taux FP. Les arbres de décision sont bien connus et des structures simples à comprendre à partir de cette règle peuvent être dérivées. L'efficacité du modèle étudié est démontrée par des résultats expérimentaux. Pour la détection du cancer du sein, l'efficacité de l'approche de l'arbre de décision a été évaluée et explorée. Tout au long de la phase de mise en œuvre, seules les valeurs numériques des caractéristiques particulières du cancer du sein sont évaluées. Les résultats expérimentaux révèlent que le classificateur d'arbre de décision a un taux d'exactitude de 100 %, tandis que la méthode Random Forest n'a qu'un taux d'exactitude de 50 %. Les performances de l'arbre de décision sont supérieures à celles de l'autre méthode pour l'ensemble de données spécifié sur la base des résultats de catégorisation des quatre algorithmes.

Les objectifs de cette recherche [328] étaient de ; i) déterminer les étapes de l'intégration CNN-XGBoost dans le diagnostic du cancer du sein et ii) calculer la précision de l'intégration CNN-XGBoost dans la détection du cancer du sein. En combinant l'apprentissage par transfert et l'augmentation des données, CNN avec XGBoost comme classificateur a été utilisé. Après avoir acquis des résultats de précision grâce à l'apprentissage par transfert, cette recherche connecte la couche finale au classificateur XGBoost. De plus, la conception de l'interface pour le processus d'évaluation a été établie à l'aide du langage de programmation Python et de la plateforme Django. Les résultats : i) les étapes d'intégration CNN-XGBoost sur des images d'histopathologie pour la détection du cancer du sein ont été découvertes. ii) atteint un niveau de précision plus élevé grâce à l'intégration CNNXGBoost pour la détection du cancer du sein. En conclusion, la détection du cancer du sein a été révélée grâce à l'intégration de CNN-XGBoost à travers des images histopathologiques. La combinaison de CNN et XGBoost peut améliorer la précision de la détection du cancer du sein.

Dans cette recherche [329], les auteurs ont utilisé les 30 caractéristiques pour extraire et prédire une prédiction précise sur le cancer du sein en utilisant une approche d'ensemble d'algorithmes d'apprentissage automatique supervisé. C'est un grand défi de concevoir un modèle d'apprentissage automatique pour évaluer les performances de la classification des tumeurs du sein. La mise en œuvre d'une méthodologie de classification efficace aidera à résoudre les complications dans l'analyse du cancer du sein. Ce modèle proposé utilise quatre algorithmes d'apprentissage automatique, des classificateurs d'arbres de décision, Random Forest KNN, une machine à vecteurs de support et une machine à vecteurs de support trouvés qui ont donné la haute précision de 0,976688 parmi eux pour la catégorisation des tumeurs du sein chez les femmes. Cette classification comprend les deux niveaux de maladie bénigne ou maligne. Le chercheur a également utilisé les autres paramètres et évalué ce modèle prédictif à l'aide de la précision, rappel et F1-Score. Le rapport d'analyse des données prouve que ce modèle prédictif a un niveau d'exactitude de 98% pour prédire le cancer à un stade précoce chez les femmes.

Dans cet article [330], les auteurs ont comparé cinq techniques d'apprentissage automatique supervisées nommées ; machine à vecteurs de support (SVM), K-plus proches voisins, forêts aléatoires, réseaux de neurones artificiels (ANN) et régression logistique. L'ensemble de données sur le cancer du sein du Wisconsin est obtenu à partir d'une importante base de données d'apprentissage automatique appelée base de données d'apprentissage automatique UCI. La performance de l'étude est mesurée par rapport à l'exactitude, la sensibilité, la spécificité, la précision, la valeur prédictive négative, le taux de faux négatifs, le taux de faux positifs, le score F1 et le coefficient de corrélation de Matthews. De plus, ces techniques ont été évaluées sur la zone de précision-rappel sous la courbe et la courbe caractéristique de fonctionnement du récepteur. Les résultats révèlent que l'ANN a obtenu l'exactitude, la précision et le score F1 les plus élevés de 98,57 %, 97,82 % et 0,9890, respectivement, tandis que l'exactitude, la précision et le score F1 de 97,14 %, 95,65 % et 0,9777 sont obtenus par SVM, respectivement.

V.3. Méthodologie et techniques

L'objectif de cette contribution est d'évaluer l'exactitude de prédiction des algorithmes de classification en termes d'efficience et d'efficacité. Dans cette section, nous avons intégré le modèle ontologique pour la prédiction du cancer de sein, pour cela nous avons comparé sept approches de classification importantes. Les approches et les matériaux utilisés, ainsi que la méthodologie expérimentale, la description de l'ensemble de données, les algorithmes d'apprentissage automatique, le modèle d'ontologie et les mesures d'évaluation, sont tous inclus dans cette section.

V.3.1. Collecte et prétraitement de l'ensemble de donnée

L'ensemble de données sur le cancer du sein a été extrait du référentiel d'apprentissage automatique de l'UCI[13]. Il y a 683 cas dans cet ensemble de données, où les cas sont bénins ou malins. La classe dans l'ensemble de données est partitionnée en 2 ou 4, où 2 correspond au cas bénin et 4 correspond au cas malin. L'ensemble de données comprend 10 attributs, neuf attributs nommés : touffe, ucz, ucp, adhérence, épithélial, nuclei, fade_chromatin, normal_nucleoli, et le dernier attribut est la classe : résultat du diagnostic (2-bénigne et 4-maligne). Il n'y avait pas d'entrées nulles dans l'ensemble de données. Une description complète de tous les attributs de l'ensemble de données est fournie dans la Table V.1.

La phase de prétraitement des données vise à préparer les données à utiliser dans le modèle de prédiction. Habituellement, les données sont désordonnées et proviennent de différentes sources avec différentes tailles et résolutions. Cette phase est donc cruciale pour nettoyer et normaliser les données afin de réduire la complexité et d'augmenter la précision du modèle de prédiction. Différents types de transformations peuvent être exécutés en fonction de l'ensemble de données, comme le redimensionnement, la rotation, le décalage, la normalisation, etc.

Nous avons trouvé que l'ensemble de données utilisé contient 234 instances similaires, après suppression des instances en double, il reste 449 instances, où 213 représentent les cas bénins et 236 représentent les cas malins.

[13] https://www.kaggle.com/datasets/ninjacoding/breast-cancer-wisconsin-benign-or-malignant

Dans l'étape de prétraitement, il est essentiel d'éliminer les valeurs manquantes, le bruit et autres anomalies dans les données sélectionnées. Toute incohérence dans les données choisies, en particulier les données liées à la maladie, peut entraîner des résultats non fiables ou un diagnostic erroné des données de test, ce qui pourrait être fatal si le modèle est mis en œuvre dans des situations réelles. L'une des étapes du prétraitement est l'élimination des variables non liées, car ces variables ne sont pas nécessaires pour atteindre l'objectif de l'étude. En outre, des valeurs manquantes ou des anomalies se produisent en raison d'un manque d'informations et de valeurs de mesure imprécises, entraînant une précision insuffisante et un pourcentage d'erreur plus élevé dans le processus d'évaluation des données.

Table V.1 - Description des attributs de l'ensemble de donnée.

Attribut	***Description***
1- clump	Épaisseur du bloc : les cellules bénignes forment souvent des monocouches, tandis que les cellules malignes forment fréquemment des multicouches.
2- ucz	Uniformité de la taille des cellules : Les cellules cancéreuses diffèrent en taille.
3- ucp	Uniformité de la forme des cellules : les cellules cancéreuses diffèrent par leur forme.
4- adhesion	Adhérence marginale : la perte d'adhérence est une indication de cancer.
5- epithelial	Taille de cellule épithéliale unique : est liée à l'uniformité mentionnée précédemment. Les cellules épithéliales considérablement développées peuvent être des cellules cancéreuses
6- bare_nuclei	Noyaux nus : Ceux-ci sont courants dans les tumeurs bénignes.
7- bland_chromatin	Dans les cellules bénignes, le noyau a une texture homogène.
8- normal_nucleoli	Dans les cellules normales, le nucléole est généralement assez petit, voire pas du tout détectable. Les nucléoles deviennent plus visibles dans les cellules cancéreuses.
9- mitoses	-
10- Class	Classe prédite (2 pour bénigne, 4 pour maligne).

V.3.2. Méthodologie de classification

L'apprentissage automatique est une approche automatisée pour apprendre où les algorithmes sont programmés pour acquérir de l'expérience à partir d'ensembles de données passés pour prédire l'avenir. Dans cette étude nous avons utilisé les algorithmes d'apprentissage automatique suivants :

- Régression logistique.
- Machine à vecteurs de support
- Forêt aléatoire.
- Naïve Bayes.
- Arbre de décision.
- K-plus proches voisins.

- Réseau de neurones artificiels.

Nous avons utilisé le logiciel Weka pour tous les algorithmes d'apprentissage automatique afin de prédire si les cellules cancéreuses sont bénignes ou malignes. Weka comprend des outils pour la classification des données, le regroupement, la visualisation, la préparation, l'exploration des règles d'association et la régression.

De plus, nous avons utilisé deux modes d'options de test : validation croisée 10 fois et répartition en pourcentage. Concernant la création du modèle ontologique, nous avons choisi l'algorithme d'arbre de décision pour implémenter les règles générées dans le modèle d'ontologie (plus de détail dans la section suivante). Nous avons choisi l'algorithme d'arbre de décision pour de nombreuses raisons, le résultat de l'arbre de classification est plus facile à comprendre et à interpréter, et il prend en charge plusieurs types de données tels que numériques, nominaux, catégoriels, etc. L'objectif de l'utilisation d'un arbre de décision est de construire un modèle d'entraînement qui peut prédire la classe ou la valeur de la variable cible en apprenant des règles de décision de base à partir de données passées (données d'entraînement).

Le résultat de la classification de l'arbre de décision ainsi un extrait de la sortie de l'arbre de décision (nous avons obtenu 11 feuilles) qui sera utilisé pour générer des règles SWRL qui seront utilisées dans le modèle d'ontologie, sont illustrés dans la Figure V.1, Figure V.2.

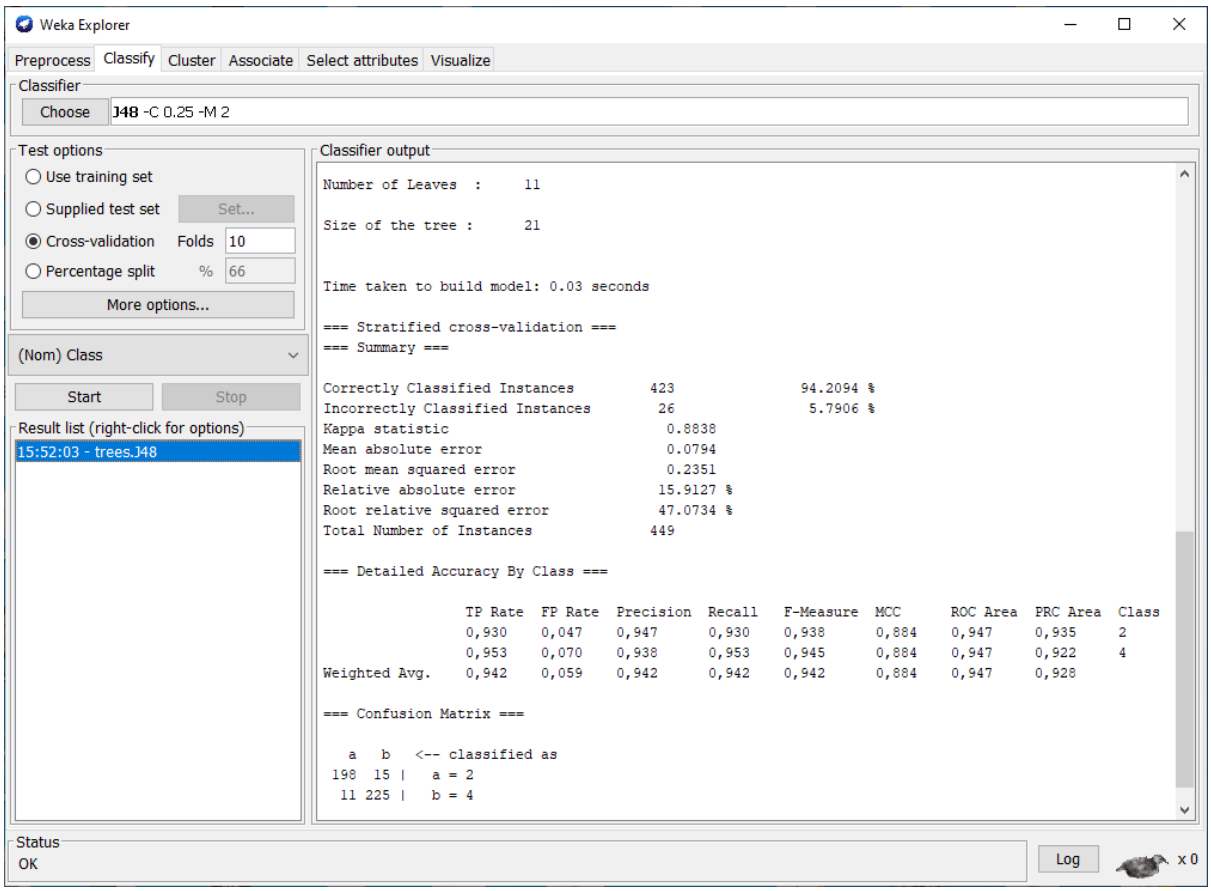

Figure V.1 - Résultat de l'arbre de décision utilisant Weka.

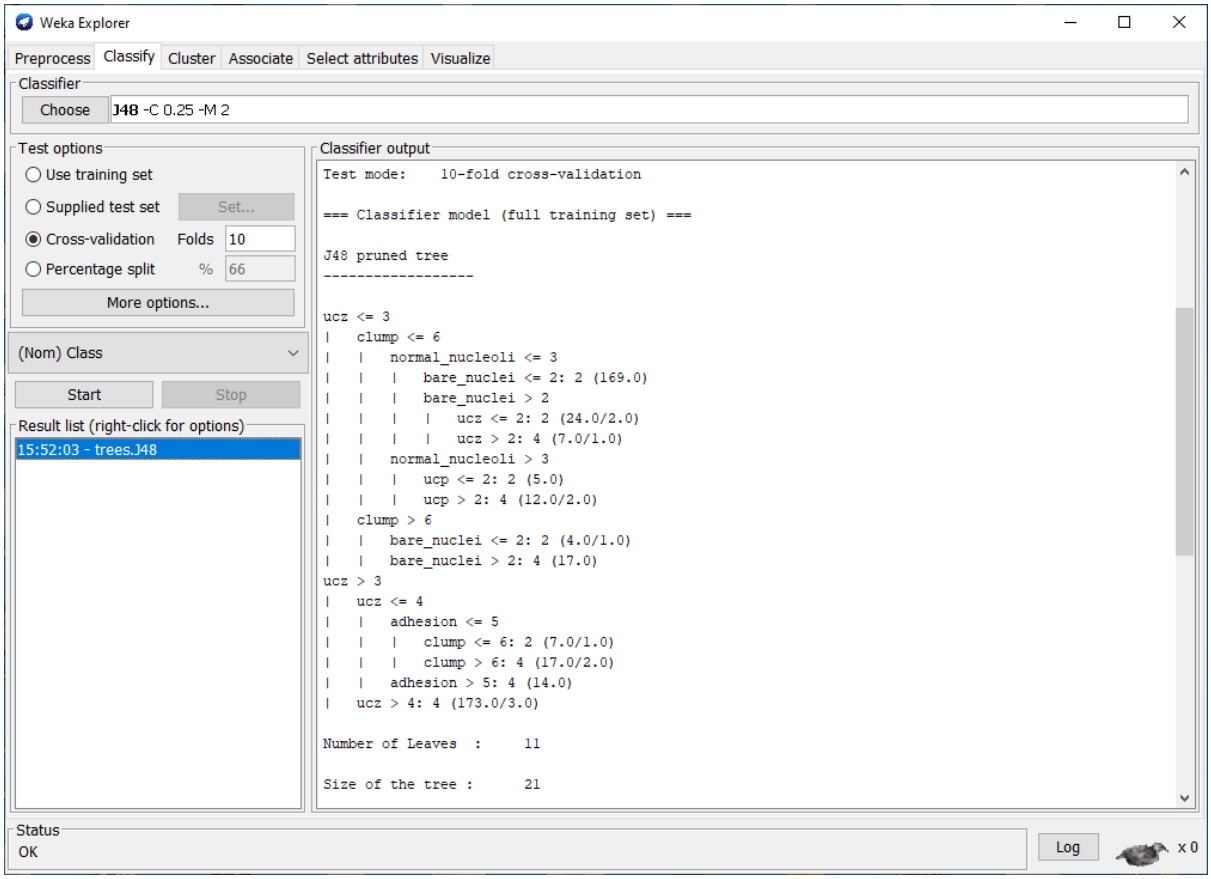

Figure V.2 - Extrait de l'arbre de décision utilisant Weka.

V.3.3. Construction de l'ontologie et raisonnement ontologique

Cette section présente les technologies utilisées pour créer l'ontologie, ainsi que l'approche utilisée pour construire le modèle d'ontologie à l'aide de règles extraites de l'algorithme d'arbre de décision.

L'ontologie a été construite à l'aide du logiciel Protégé, une plate-forme open source qui fournit un ensemble d'outils à une communauté d'utilisateurs croissante pour construire des modèles de domaine et des applications basées sur les connaissances avec des ontologies. L'ontologie a été créée manuellement ; les classes principales sont Diagnostic et Patient. La représentation graphique de l'ontologie est illustrée dans la Figure V.3.

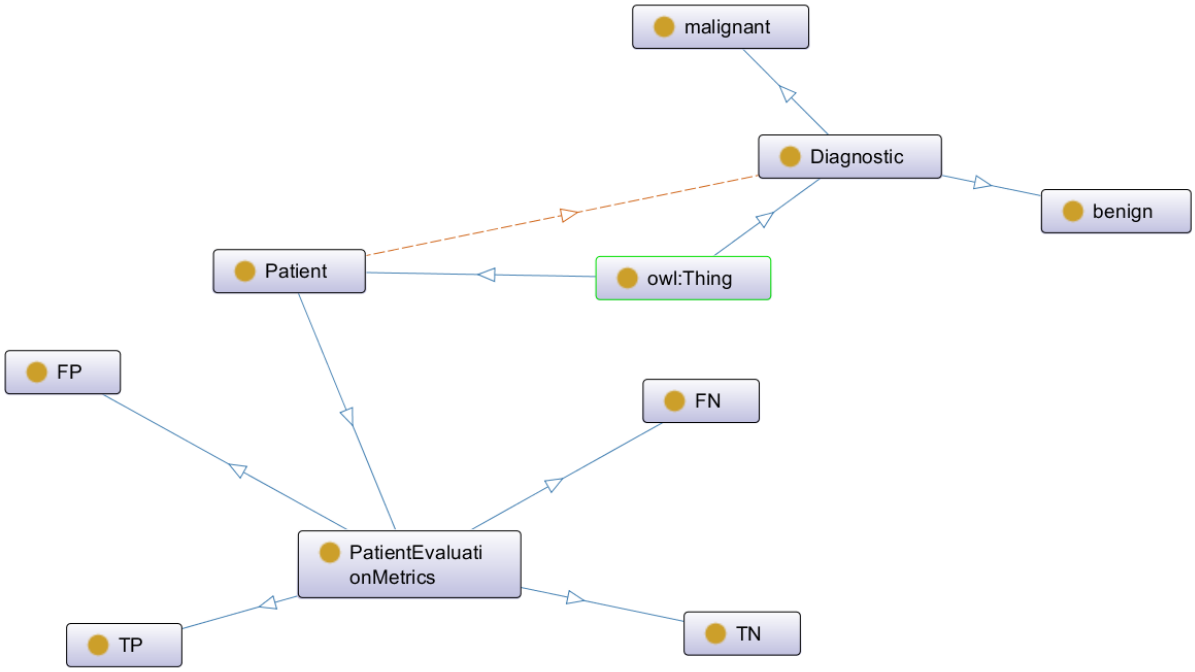

Figure V.3 - Représentation graphique de l'ontologie.

Les propriétés de données utilisées dans l'ontologie sont les mêmes attributs présentés dans la Table V.1 qui sont utilisés pour construire des modèles d'algorithmes d'apprentissage automatique. La Figure V.4 illustre les propriétés des données. Un plugin parmi les plugins logiciels Protégé appelé Cellfie est utilisé pour importer le même ensemble de données « Breast Cancer WISCONSIN » utilisé dans Weka.

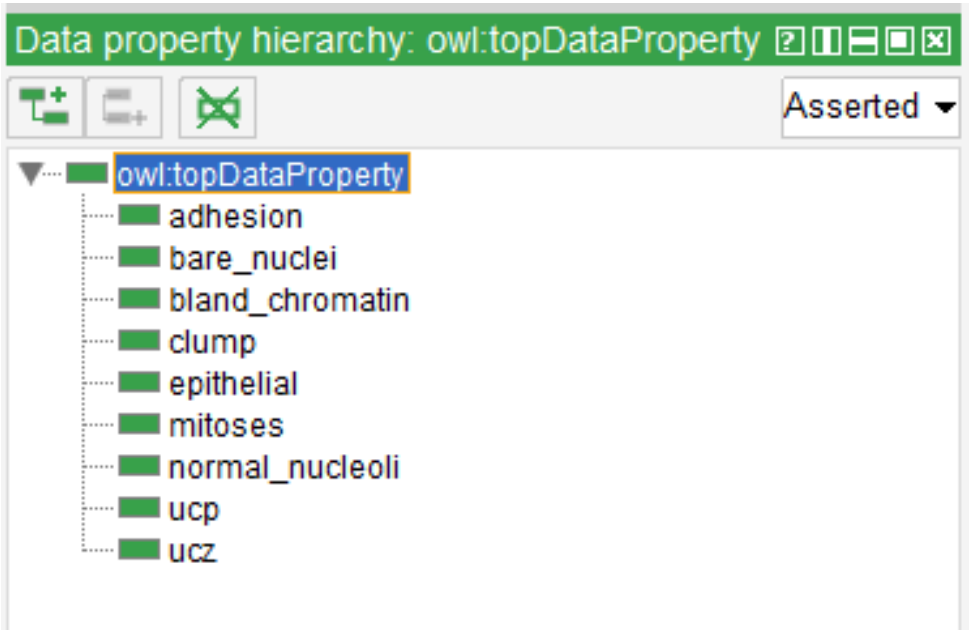

Figure V.4 - Propriétés des données de l'ontologie.

Suite à la création de classes, de propriétés de données et d'instances dans l'ontologie. Nous devons établir les règles de raisonnement SWRL. Pour ce faire, nous avons utilisé le plugin SWRLTab, nous avons récupéré les règles créées à partir de l'algorithme de l'arbre de

décision et les avons importées dans Protégé. Les règles collectées à partir de l'algorithme d'arbre de décision sont converties à l'aide du langage de programmation Java, chaque feuille de l'arbre étant extraite en tant que règle SWRL unique. Par exemple :

Une feuille de l'algorithme de l'arbre de décision :

If ucp > 2 && ucz ≤ 4 && bare_nuclei ≤ 2 && adhesion ≤ 3 THEN put the patient in benign

SWRL obtenu :

Patient(?P) ^ ucp(?P, ?UCP) ^ swrlb:greaterThan(?UCP, '2'^^xsd:decimal) ^ ucz(?P, ?UCZ) ^ swrlb:lessThanOrEqual(?UCZ, '4'^^xsd:decimal) ^ bare_nuclei(?P, ?BN) ^ swrlb:lessThanOrEqual(?BN, '2'^^xsd:decimal) ^ adhesion(?P, ?A) ^ swrlb:lessThanOrEqual(?A, '3'^^xsd:decimal) → benign

Pour exécuter les règles SWRL et déduire de nouveaux axiomes d'ontologie, nous avons utilisé un autre plugin du logiciel Protégé nommé Pellet, qui inclut des capacités pour vérifier la cohérence de l'ontologie, traiter les règles SWRL, calculer la hiérarchie de classification, traiter OWL, expliquer les inférences et répondre aux requêtes SPARQL. Il utilise les règles Ontologie et SWRL pour initier l'inférence, puis détermine si les cellules cancéreuses sont bénignes ou malignes. Les résultats du classificateur d'ontologie sont rapportés dans la section suivante.

V.3.4. Évaluation

Afin d'évaluer les performances des modèles, différents scores ont été mesurés pour garantir les résultats tels que l'exactitude, la précision, le rappel, la F-mesure. Nous avons utilisé deux modes de test (split-test et validation croisée K-fold) pour analyser nos résultats expérimentaux. De plus, les mêmes critères sont utilisés pour évaluer la validité de cette recherche comparative, y compris les classificateurs d'apprentissage automatique et le modèle ontologique.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$\text{F-Measure} = 2 * \frac{PREC * REC}{PREC + REC}$$

Il est important de mentionner que toutes ces mesures ont été calculées pour tous les classificateurs différents dans toutes les expériences réalisées.

V.4. Mise en œuvre et analyse des résultats

Dans cette section, après avoir implémenté les algorithmes d'apprentissage automatique et le modèle ontologique, nous avons analysé leurs performances sur l'ensemble de données. Ceci

est effectué en appliquant une technique de validation croisée à dix volets, c'est-à-dire que l'ensemble de données faisait partie de dix portions. La technique de validation croisée en dix volets est utilisée pour valider le modèle délibéré. Dans cette technique, neuf fois est utilisé pour l'entraînement et le reste pour les tests. Nous avons également appliqué un fractionnement en pourcentage (split 50% train, remainder test) dans le but d'enrichir l'étude.

La matrice de confusion est calculée pour chaque algorithme incluant le modèle d'ontologie. Il est généré pour le résultat réel et prédit composé de TP, FP, TN et FN pour calculer l'exactitude, la précision, le rappel et la F-mesure pour chaque algorithme utilisé. Ci-dessous, la signification des termes est mentionnée :

- TP = vrai positif (identifié avec précision).
- TN = vrai négatif (identifié de manière inexacte).
- FP = faux positif (rejeté avec précision).
- FN = faux négatif (rejeté de manière inexacte).

Les résultats de l'évaluation des différents classificateurs qui ont été utilisés dans cette étude sont présentés aux figures : Figure V.6, Figure V.7, Figure V.8, Figure V.9. Les statistiques et les résultats du modèle ontologique sont également présentés dans les tableaux : Table V.2, Table V.3 et la Figure V.5 illustre les mesures de performance du modèle ontologique. La Table V.4 résume également les résultats des différents classificateurs qui ont été utilisés dans cette recherche.

- **Exactitude :**

Selon la Figure V.6 et la Table V.4, le modèle ontologique a atteint la valeur maximale de 96,88 % et la forêt aléatoire avec un taux de 96,00 %, et 95,30 % pour la machine à vecteur de support et la régression logistique en termes de validation croisée de 10 fois. Presque les mêmes résultats en utilisant le mode de test fractionné, nous avons obtenu 96,00 %, 95,10 % pour l'ontologie et la forêt aléatoire consécutivement, et 94,60 % pour la machine à vecteur de support et le réseau de neurones artificiels.

- **Précision :**

Le classificateur d'ontologie a la précision la plus élevée de 97,64 % en termes de mode de validation croisée 10 fois, suivi par Random Forest et Naïve Bayes. En ce qui concerne le mode de test fractionné, la valeur de précision la plus élevée de 97,00 % va pour ANN. Plus de détails sont présentés dans la Table V.4 et la Figure V.7.

- **Rappel :**

Selon la Figure V.8 et la Table V.4, le modèle ontologique a les valeurs de rappel les plus élevées de 95,83 % et 97,00 % pour les deux modes de test, suivis par Random Forest, Logistic Regression et KNN pour le mode de validation croisée 10 fois, et arbre de décision et forêt aléatoire pour le mode de test fractionné.

- **F-mesure :**

Selon la Figure V.9 et la Table V.4, le modèle d'ontologie avait la plus grande valeur de 96 % dans les deux modes de test, suivi de Random Forest en deuxième position et de Support Vector Machine en troisième position.

Table V.2 - Validation croisée 10 fois pour le modèle ontologique.

Matrice de confusion		**Classe réelle**	
		positive	**négative**
Classe prédite	**positive**	TP : 207	FP : 5
	négative	FN : 9	TN : 228

Table V.3 - Mode split 50 % pour le modèle ontologique.

Matrice de confusion		**Classe réelle**	
		positive	**négative**
Classe prédite	**positive**	TP : 98	FP : 6
	négative	FN : 3	TN : 117

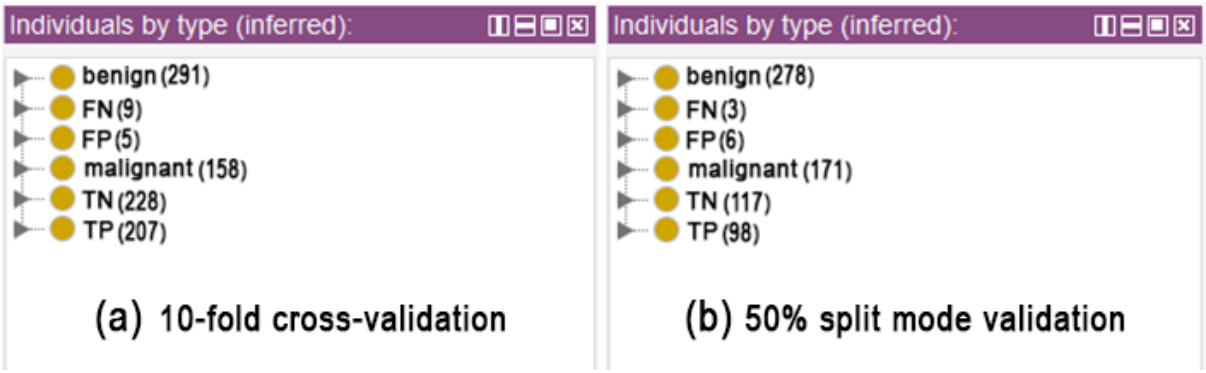

Figure V.5 - Résultats des concepts inférés.

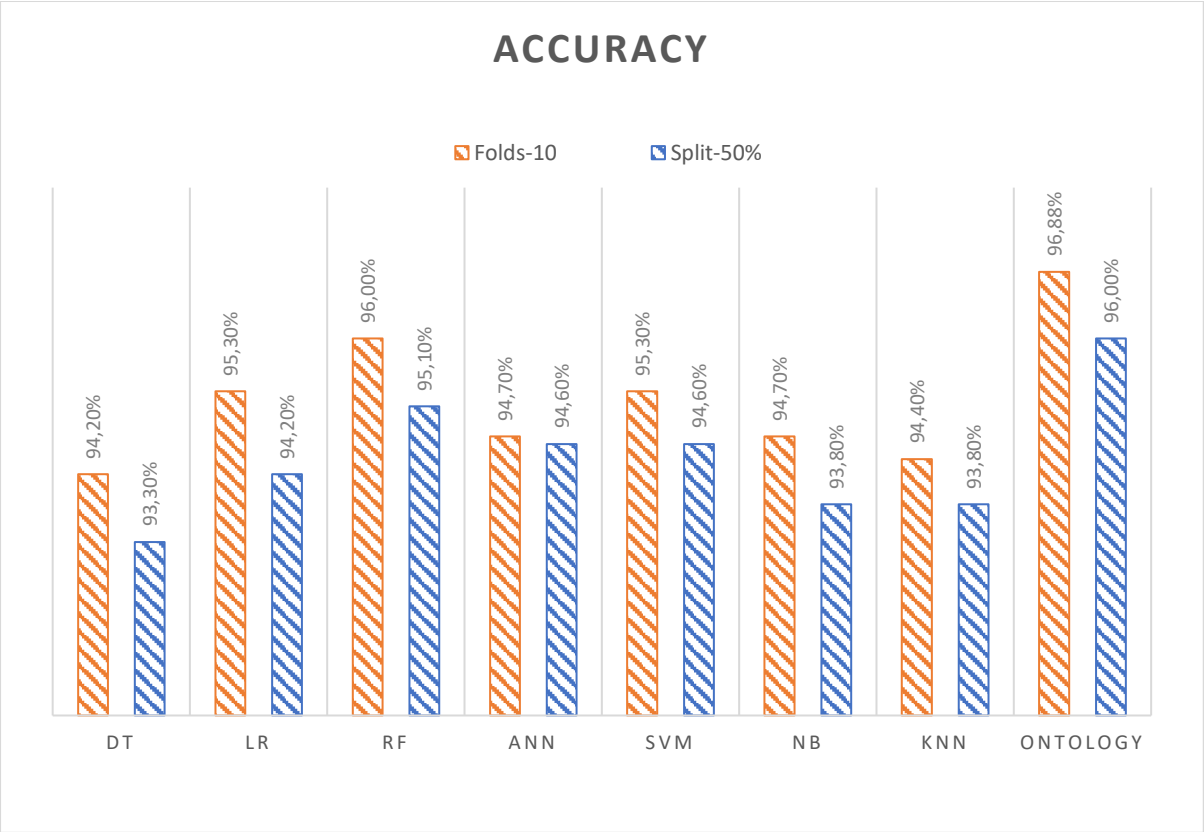

Figure V.6 - Résultats de comparaison de l'exactitude.

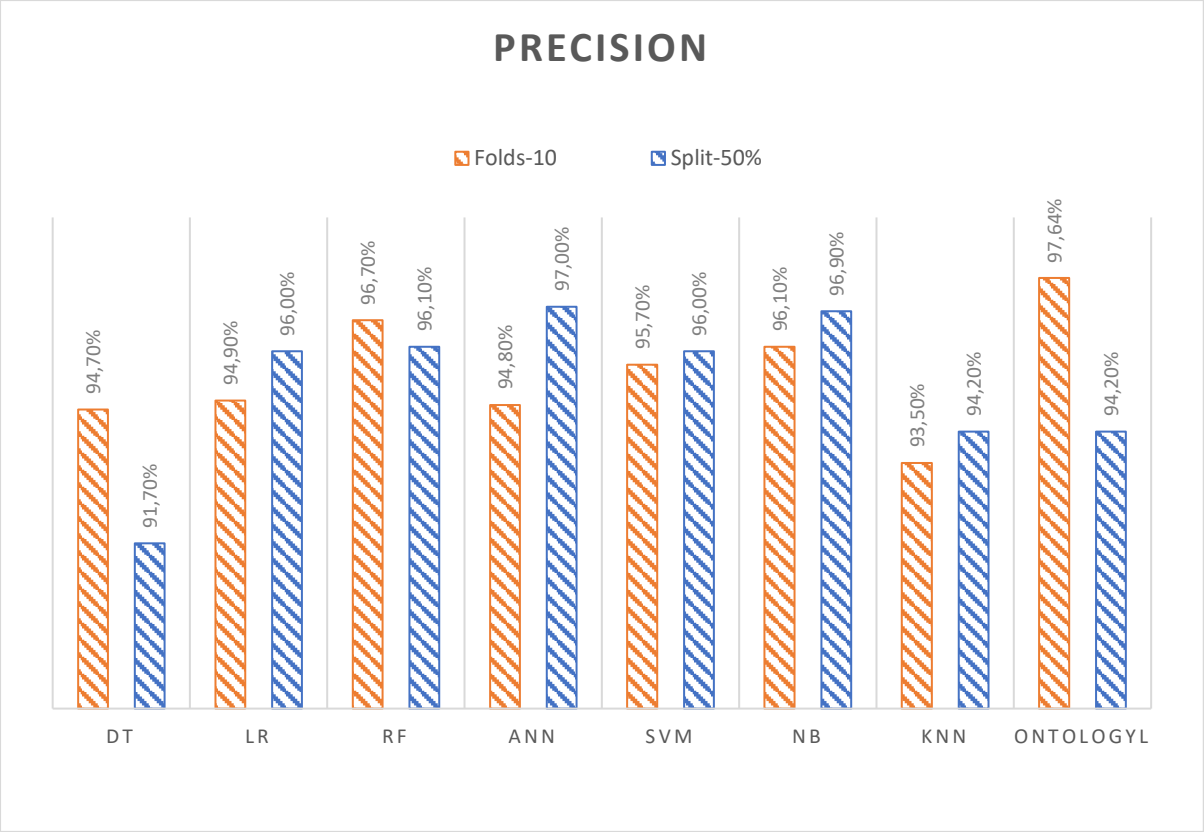

Figure V.7 - Comparaison des résultats de précision.

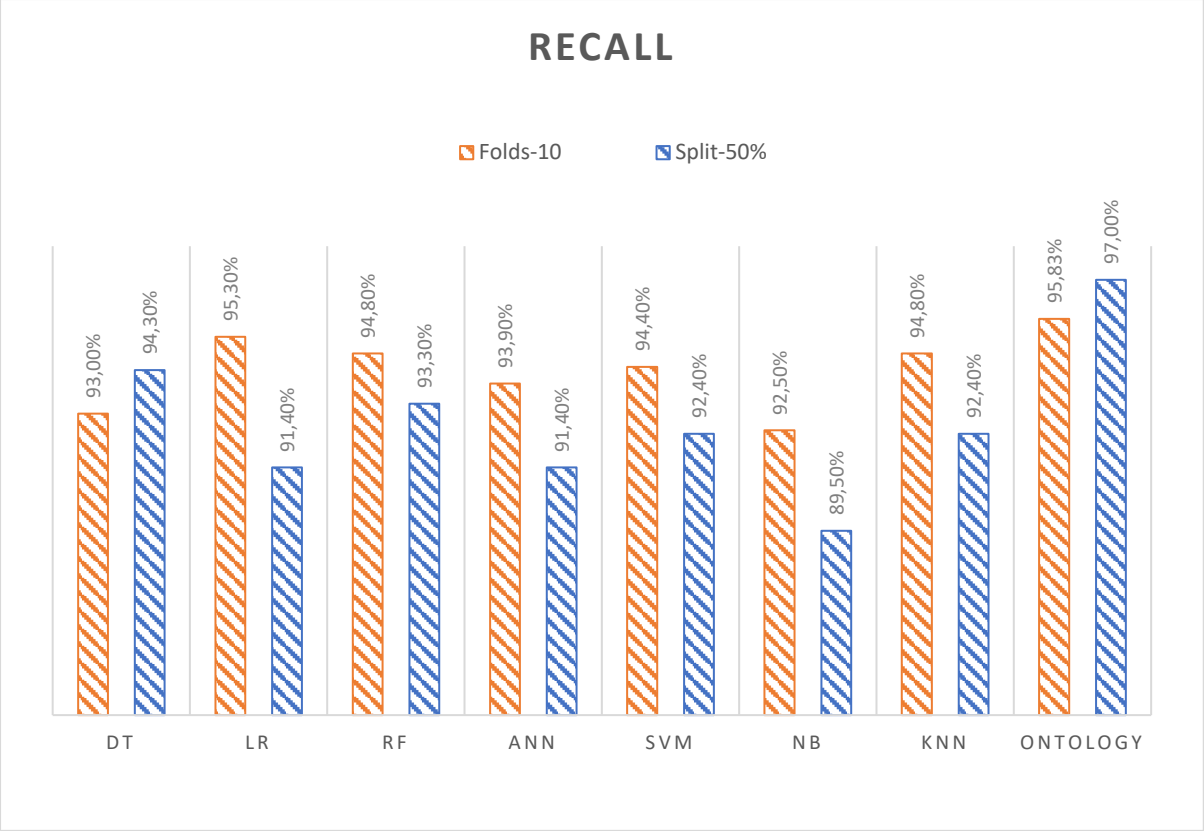

Figure V.8 - Comparaison des résultats de rappel.

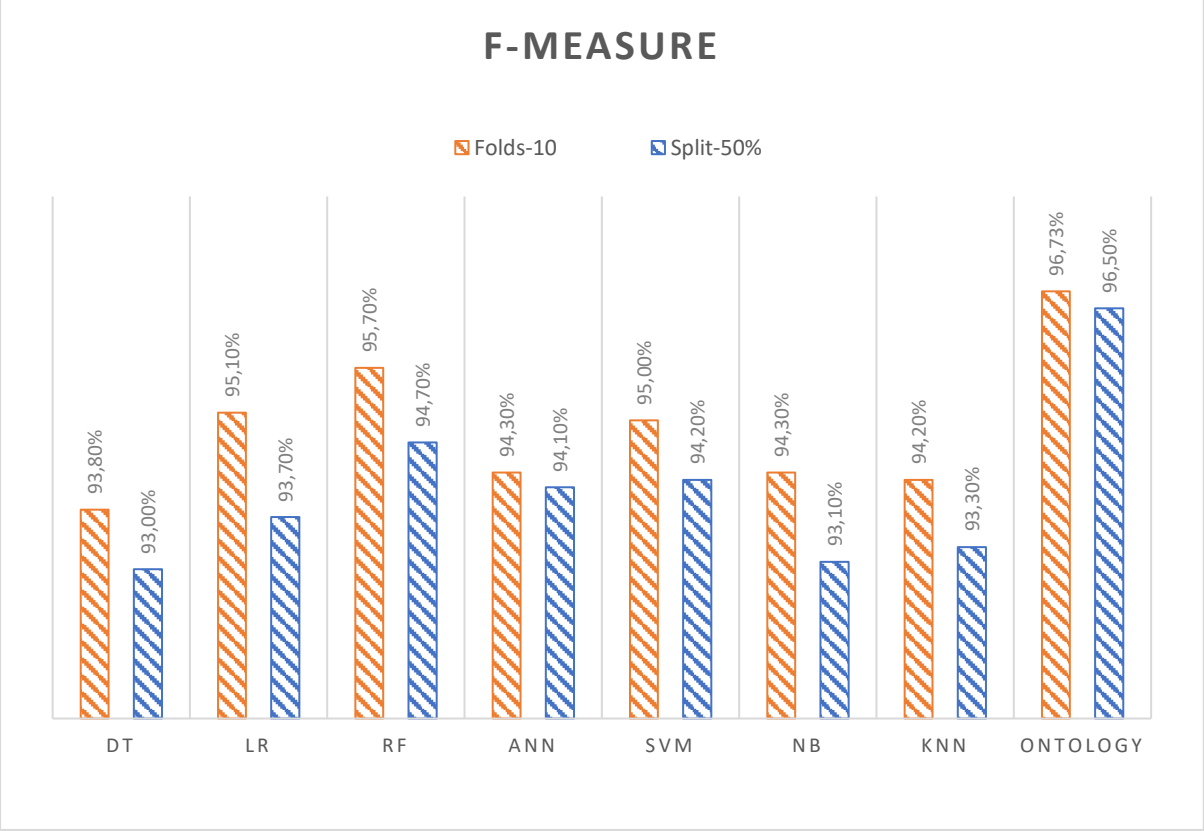

Figure V.9 - Résultats de la comparaison F-Measure.

Les résultats expérimentaux révèlent que le modèle d'ontologie a la précision la plus élevée de 96,9 %, suivi de la forêt aléatoire à 96,00 % et à la fois de la régression logistique et de la machine à vecteurs de support à 95,30 %. En termes de données indiquées ci-dessus, nous ne voyons aucune différence significative entre les modes de test 50%-Split et 10-Folds. Nous concluons que le modèle ontologique peut aider en étendant la portée du modèle d'apprentissage automatique. La combinaison du modèle ontologique avec l'apprentissage automatique peut donner de bons résultats. Le modèle ontologique obtient des résultats comparables aux classificateurs d'apprentissage automatique. De plus, l'ontologie correspond à l'objectif de toute organisation, qui peut être mathématique, logique ou sémantique.

Table V.4 - Résultats du modèle ontologique et des classificateurs d'apprentissage automatique.

	Exactitude		**Précision**		**Rappel**		**F-mesure**	
	Folds-10	Split-50%	Folds-10	Split-50%	Folds-10	Split-50%	Folds-10	Split-50%
Decision Tree	0.942	0.933	0.947	0.917	0.93	0.943	0.938	0.93
Logistic Regression	0.953	0.942	0.949	0.96	0.953	0.914	0.951	0.937
Random Forest	0.96	0.951	0.967	0.961	0.948	0.933	0.957	0.947
Artificial Neural Network	0.947	0.946	0.948	0.97	0.939	0.914	0.943	0.941
Support Vector Machine	0.953	0.946	0.957	0.96	0.944	0.924	0.95	0.942
Naïve Bayes	0.947	0.938	0.961	0.969	0.925	0.895	0.943	0.931
K-Nearest Neighbors	0.944	0.938	0.935	0.942	0.948	0.924	0.942	0.933
Ontology Model	0.969	0.960	0.976	0.942	0.958	0.970	0.967	0.965

V.5. Conclusion

La détection précoce de la maladie est devenue un problème crucial en raison de la croissance rapide de la population dans la recherche médicale ces derniers temps. Avec la croissance rapide de la population, le risque de décès lié au cancer du sein augmente de façon exponentielle. Le cancer du sein est le deuxième cancer le plus grave parmi tous les cancers déjà dévoilés. Un système de détection automatique des maladies aide le personnel médical à diagnostiquer les maladies et offre une réponse fiable, efficace et rapide tout en réduisant le risque de décès. Dans cet article, nous avons intégré un modèle d'ontologie avec l'algorithme d'arbre de décision en comparaison avec sept techniques d'apprentissage automatique supervisées nommées support vector machine (SVM), K-plus proches voisins, forêts aléatoires, réseaux de neurones artificiels (ANN) et régression logistique. L'ensemble de données sur le cancer du sein du Wisconsin est obtenu à partir d'une importante base de données d'apprentissage automatique appelée base de données d'apprentissage automatique UCI. La performance de l'étude est mesurée en termes d'exactitude, de précision, de rappel et de F-mesure. Les résultats expérimentaux révèlent que le modèle d'ontologie a la précision la plus élevée de 96,9 %, suivi de la forêt aléatoire à 96,00 % et à la fois de la régression logistique et de la machine à vecteurs de support à 95,30 %. La combinaison du modèle ontologique avec l'apprentissage automatique peut donner de bons résultats. Le modèle ontologique obtient des résultats comparables aux classificateurs d'apprentissage automatique. Nous concluons que le modèle ontologique peut aider en étendant la portée du modèle d'apprentissage automatique. Dans des travaux futurs, nous souhaitons améliorer cette analyse comparative en adoptant de nouvelles façons d'incorporer des règles d'apprentissage automatique avec la méthode des modèles ontologiques.

CONCLUSION GENERALE

La science des données est un domaine multidisciplinaire qui utilise des méthodes, des processus, des algorithmes et des systèmes scientifiques pour extraire des connaissances et des idées à partir de données structurées et non structurées. Les statistiques, l'exploration de données, la visualisation de données, l'apprentissage automatique, l'apprentissage en profondeur et l'intelligence artificielle sont les principaux sous-thèmes de la science des données. Même si la science des données est née dans les années 1990, l'importance de ce domaine se réalise de nos jours. Il est mentionné dans différentes études que la quantité de données dans le monde augmente rapidement et que le type de données non structurées représente toujours plus de la moitié de la quantité totale de données. Par conséquent, la science des données est devenue un enjeu essentiel dans tous les domaines pour rendre les données compréhensibles. La santé est l'un des environnements nécessaires aux applications de la science des données puisque le Big Data en fait partie. Le volume de données collectées dans le domaine de la santé est énorme, pourtant il est prouvé que 80% des données collectées ne sont pas organisées. Le nombre total d'études parmi les applications de la science des données dans le domaine de la santé a considérablement augmenté.

L'environnement de la santé est l'un des domaines les plus précis pour les applications de science des données en raison de la quantité de données qu'il contient et de la pertinence du type de données. Le flux de données dans les hôpitaux est un processus continu et comprend des valeurs numériques en général. Healthcare est un système ouvert d'amélioration avec des études sur l'exploration de données et les techniques d'apprentissage automatique. Les chercheurs affirment que l'expertise sur un ordinateur vous donnerait des résultats significatifs et la possibilité de prédire l'avenir avec l'historique des données passées. De nombreuses études ont été réalisées sur des ensembles de données sur des maladies différentes, et la plupart d'entre elles ont une précision de classification suffisante.

Les techniques d'apprentissage automatique sont largement utilisées dans toutes les disciplines scientifiques et ont révolutionné les industries du monde entier. L'application d'outils et d'algorithmes d'apprentissage automatique dans les soins de santé a récemment connu des progrès significatifs. Ces procédés ont démontré leur efficacité et peuvent être bénéfiques dans le traitement de maladies chroniques telles que les maladies cardiovasculaires. De plus, le Web sémantique, pour sa part, a démontré sa valeur et sa force dans diverses disciplines, dont la santé. L'ontologie, en tant que composant du Web sémantique, a la capacité de traiter les concepts et les relations de la même manière que les humains voient les concepts connectés.

Dans le domaine médical, les ontologies ont été appliquées non seulement pour la modélisation et l'analyse des données mais également pour la classification des données. En effet, elles sont liées à plusieurs techniques de classification à savoir l'apprentissage automatique ou le deep learning (apprentissage profond). Le fait de combiner les ontologies aux techniques issues des domaines de l'apprentissage automatique caractérise un domaine récent permettant de concevoir de nouvelles stratégies de raisonnement. De plus, l'ontologie a été l'une des techniques les plus largement utilisées pour gérer, organiser et extraire des données au cours des dernières décennies. C'est un mode de représentation des données qui a été efficacement utilisé dans un certain nombre de domaines, en particulier le domaine médical. Il est important en informatique en raison de sa capacité à exprimer de nombreux

concepts et leurs relations entre les domaines. En réalité, aucune ontologie unique n'est suffisante pour répondre aux demandes croissantes de soins de santé d'aujourd'hui, et les ontologies doivent être combinées avec des algorithmes d'apprentissage automatique pour faciliter l'intégration et l'analyse des données.

La détection précoce de la maladie est devenue un problème crucial en raison de la croissance rapide de la population dans la recherche médicale ces derniers temps. Avec la croissance rapide de la population, le risque de décès lié aux maladies augmente de façon exponentielle. Un système de détection automatique des maladies qui aide le personnel médical à diagnostiquer les maladies et offre une réponse fiable, efficace et rapide tout en réduisant le risque de décès est nécessaire.

Nous avons introduit une nouvelle approche consistant à fusionner l'apprentissage automatique et le Web Sémantique. D'une part, les algorithmes d'apprentissage automatique apprennent de manière autonome à effectuer une tâche ou à faire des prédictions à partir de données et améliorent leurs performances dans le temps, tandis que le Web sémantique fournit plusieurs formats d'affichage des données et des connaissances ontologiques de base. La fusion des deux nous a permis de construire un modèle basé sur une ontologie capable de prédire les maladies avec une grande précision. Nous avons établi un modèle ontologique de représentation des connaissances à travers le langage OWL, les règles SWRL et le raisonneur. Pour ce faire, nous générons les règles à partir de l'algorithme d'arbre de décision, puis nous les implémentons dans l'ontologie en utilisant le langage de règles pour le web sémantique (SWRL). Nous présentons également une analyse comparative entre les sept techniques de classification populaires (arbre de décision, algorithme de forêt aléatoire, méthode des k plus proches voisins, Naïve Bayes, machine à vecteurs de support, régression logistique et réseau de neurones artificiels) et la classification d'apprentissage automatique basée sur l'ontologie en utilisant des paramètres soigneusement choisis tels que la Précision, l'Exactitude, le Rappel et la F-mesure, qui sont dérivés de la matrice de confusion.

En plus d'identifier le meilleur modèle de classificateur qui introduit une plus grande précision de classification pour les ensembles de données prédéfini utilisé dans notre étude, le processus de classification des données est mis en œuvre en appliquant des opérations de prétraitement et en extrayant des caractéristiques aux enregistrements de données spécifiés à partir de l'ensemble de données à l'aide de WEKA. Le WEKA (Waikato Environment for Knowledge Analysis) est un logiciel open-source qui contient un ensemble d'algorithmes pour les tâches d'exploration de données. Ces algorithmes peuvent être appliqués à un ensemble de données soit directement via l'interface WEKA, soit via du code Java. Ensuite, les différents classificateurs sont implémentés avec différentes variables en utilisant plusieurs algorithmes et plusieurs options pour calculer le meilleur rapport de l'exactitude.

Selon les résultats, le modèle ontologique a dépassé tous les algorithmes d'apprentissage automatique dans tous les différents cas expérimentaux, avec une valeur d'exactitude la plus élevée dans les trois ensembles de données utilisées. Nous concluons que le modèle ontologique peut aider en étendant la portée du modèle d'apprentissage automatique. Ils peuvent comprendre n'importe quel type ou variation de données, et chaque donnée peut être affectée à un certain travail. La combinaison du modèle ontologique avec l'apprentissage automatique peut donner de bons résultats. Le modèle ontologique obtient des résultats comparables aux classificateurs d'apprentissage automatique. Les humains peuvent

interpréter les résultats et les règles peuvent être modifiées ou ajoutées au besoin. De plus, il prend en charge les formats de données non structurés, semi-structurés et structurés, permettant une intégration plus transparente des données. Il peut comprendre tous les aspects du processus de modélisation des données, en commençant par les schémas au niveau le plus élémentaire. En conséquence, ils peuvent gérer les quantités massives de données utilisées comme entrées pour l'entraînement à l'apprentissage automatique ou les sorties comme résultats. De plus, l'ontologie correspond à l'objectif de toute organisation, qui peut être mathématique, logique ou sémantique.

Ces contributions peuvent servir des systèmes d'aide à la décision pour les médecins, en utilisant des modèles développés comme aide pour détecter la présence de maladie chez une personne en fonction des symptômes déclarés. Ce type des contributions peuvent encourager les individus à consulter immédiatement le médecin et favorise le diagnostic précoce de la maladie. Les modèles développés peuvent être utilisé pour construire une application avec l'avantage de l'utiliser comme une évaluation préliminaire du patient pour les médecins praticiens.

Ces contributions seront continuellement améliorées dans le futur, nous prévoyons ensuite d'explorer la méthodologie de prédiction en utilisant l'ensemble de données mis à jour et d'utiliser les méthodes d'apprentissage automatique les plus précises et les plus appropriées pour la prévision. Les prévisions en temps réel seront l'un des principaux axes de nos travaux futurs. Nous souhaitons également automatiser le processus de notre approche en utilisant les APIs java des plateformes et logiciels utilisés.

LISTE DES PUBLICATIONS

Conférences

- H. El Massari, S. Mhammedi, N. Gherabi, and M. Nasri, "**Virtual OBDA Mechanism Ontop for Answering SPARQL Queries Over Couchbase**," in Advanced Technologies for Humanity, in Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing, 2022, pp. 193–205. doi: 10.1007/978-3-030-94188-8_19.
- H. El Massari, S. Mhammedi, Z. Sabouri, and N. Gherabi, "**Ontology-Based Machine Learning to Predict Diabetes Patients**," in Advances in Information, Communication and Cybersecurity, Cham, 2022, pp. 437–445. doi: 10.1007/978-3-030-91738-8_40.
- H. El Massari, N. Gherabi, S. Mhammedi, Z. Sabouri, H. Ghandi, and F. Qanouni, "**Effectiveness of applying Machine Learning techniques and Ontologies in Breast Cancer detection**," Procedia Comput. Sci., vol. 218, pp. 2392–2400, Jan. 2023, doi: 10.1016/j.procs.2023.01.214.
- **COVID-19 Prediction Applying Machine learning and ontological language** (en cours de publication).

Journaux

- H. El Massari, S. Mhammedi, and N. Gherabi, "**Bridging the gap between the semantic web and big data: answering SPARQL queries over NoSQL databases**," Int. J. Electr. Comput. Eng. IJECE, vol. 12, no. 6, Art. no. 6, Dec. 2022, doi: 10.11591/ijece.v12i6.pp6829-6835.
- H. El Massari, N. Gherabi, S. Mhammedi, Z. Sabouri, and H. Ghandi, "**Ontology-Based Decision Tree Model for Prediction of Cardiovascular Disease**," Indian J. Comput. Sci. Eng., vol. 13, no. 3, pp. 851–859, Jun. 2022, doi: 10.21817/indjcse/2022/v13i3/221303143.
- H. El Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "**Diabetes Prediction Using Machine Learning Algorithms and Ontology**," J. ICT Stand., pp. 319–338, May 2022, doi: 10.13052/jicts2245-800X.10212.
- H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, M. Bahaj, and M. R. Naqvi, "**The Impact of Ontology on the Prediction of Cardiovascular Disease Compared to Machine Learning Algorithms**," Int. J. Online Biomed. Eng. IJOE, vol. 18, no. 11, Art. no. 11, Aug. 2022, doi: 10.3991/ijoe.v18i11.32647.
- H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, F. Qanouni, and M. Bahaj, "**Integration of ontology with machine learning to predict the presence of covid-19 based on symptoms**," Bull. Electr. Eng. Inform., vol. 11, no. 5, Art. no. 5, Oct. 2022, doi: 10.11591/eei.v11i5.4392.
- H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, F. Qanouni, and M. Bahaj, "**An Ontological Model based on Machine Learning for Predicting Breast Cancer**," Int. J. Adv. Comput. Sci. Appl. IJACSA, vol. 13, no. 7, Art. no. 7, 31 2022, doi: 10.14569/IJACSA.2022.0130715.

BIBLIOGRAPHIE

[1] V. Kotu and B. Deshpande, *Data Science Concepts and Practice*. Morgan Kaufmann, 2019. Accessed: Dec. 17, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128147610000010

[2] P. Hitzler, "A review of the semantic web field," *Commun. ACM*, vol. 64, no. 2, pp. 76–83, Jan. 2021, doi: 10.1145/3397512.

[3] S. Staab and R. Studer, *Handbook on Ontologies*. in International Handbooks on Information Systems, no. 1. Springer Berlin, Heidelberg, 2004. Accessed: Dec. 17, 2022. [Online]. Available: https://link.springer.com/book/10.1007/978-3-540-24750-0

[4] C. K. Leung *et al.*, "Data science for healthcare predictive analytics," in *Proceedings of the 24th Symposium on International Database Engineering & Applications*, in IDEAS '20. New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 1–10. doi: 10.1145/3410566.3410598.

[5] V. Dhar, "Data science and prediction," *Commun. ACM*, vol. 56, no. 12, pp. 64–73, Dec. 2013, doi: 10.1145/2500499.

[6] B. Fulkerson, "Machine Learning, Neural and Statistical Classification," *Technometrics*, vol. 37, no. 4, pp. 459–459, Nov. 1995, doi: 10.1080/00401706.1995.10484383.

[7] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, Dec. 2016, pp. 1–4. doi: 10.1109/ICEDSA.2016.7818560.

[8] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis," *Cancer Inform.*, vol. 2, p. 117693510600200030, Jan. 2006, doi: 10.1177/117693510600200030.

[9] M. Nemoto *et al.*, "Machine Learning for Computer-aided Diagnosis," *Jpn. J. Med. Phys. Igakubutsuri*, vol. 36, no. 1, pp. 29–34, 2016, doi: 10.11323/jjmp.36.1_29.

[10] S. Sahran *et al.*, *Machine Learning Methods for Breast Cancer Diagnostic*. IntechOpen, 2018. doi: 10.5772/intechopen.79446.

[11] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of Breast Cancer Detection Using Different Machine Learning Techniques," in *Data Mining and Big Data*, Y. Tan, Y. Shi, and M. Tuba, Eds., in Communications in Computer and Information Science. Singapore: Springer, 2020, pp. 108–117. doi: 10.1007/978-981-15-7205-0_10.

[12] I. Mihaylov, M. Nisheva, and D. Vassilev, "Machine Learning Techniques for Survival Time Prediction in Breast Cancer," in *Artificial Intelligence: Methodology, Systems, and Applications*, G. Agre, J. van Genabith, and T. Declerck, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 186–194. doi: 10.1007/978-3-319-99344-7_17.

[13] H. Asri, H. Mousannif, and H. Al Moatassim, "A Hybrid Data Mining Classifier for Breast Cancer Prediction," in *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)*, M. Ezziyyani, Ed., in Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020, pp. 9–16. doi: 10.1007/978-3-030-36664-3_2.

[14] S. Sadhukhan, N. Upadhyay, and P. Chakraborty, "Breast Cancer Diagnosis Using Image Processing and Machine Learning," in *Emerging Technology in Modelling and Graphics*, J. K. Mandal and D. Bhattacharya, Eds., in Advances in Intelligent Systems and Computing. Singapore: Springer, 2020, pp. 113–127. doi: 10.1007/978-981-13-7403-6_12.

[15] H. Benbrahim, H. Hachimi, and A. Amine, “Comparative Study of Machine Learning Algorithms Using the Breast Cancer Dataset,” in *Advanced Intelligent Systems for Sustainable Development (AI2SD’2019)*, M. Ezziyyani, Ed., in Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020, pp. 83–91. doi: 10.1007/978-3-030-36664-3_10.

[16] A. Osmanović, S. Halilović, L. A. Ilah, A. Fojnica, and Z. Gromilić, “Machine Learning Techniques for Classification of Breast Cancer,” in *World Congress on Medical Physics and Biomedical Engineering 2018*, L. Lhotska, L. Sukupova, I. Lacković, and G. S. Ibbott, Eds., in IFMBE Proceedings. Singapore: Springer Nature, 2019, pp. 197–200. doi: 10.1007/978-981-10-9035-6_35.

[17] R. Negi and R. Mathew, “Machine Learning Algorithms for Diagnosis of Breast Cancer,” in *Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2018)*, A. P. Pandian, T. Senjyu, S. M. S. Islam, and H. Wang, Eds., in Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing, 2020, pp. 928–932. doi: 10.1007/978-3-030-24643-3_109.

[18] The author is with the Department of Electrical Engineering and Computer Science, University of Toledo, OH 43606 USA and A. A. Bataineh, “A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection,” *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 248–254, Jun. 2019, doi: 10.18178/ijmlc.2019.9.3.794.

[19] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, Jan. 2016, doi: 10.1016/j.procs.2016.04.224.

[20] L. R. Borges, “Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection,” in *Workshop de Visão Computacional*, 2015, pp. 15–19.

[21] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, “Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm,” in *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*, Aug. 2016, pp. 35–39. doi: 10.1109/DeSE.2016.8.

[22] N. Sinha, P. Sharma, and D. Arora, “Prediction Model for Breast Cancer Detection Using Machine Learning Algorithms,” in *Computational Methods and Data Engineering*, V. Singh, V. K. Asari, S. Kumar, and R. B. Patel, Eds., in Advances in Intelligent Systems and Computing. Singapore: Springer, 2021, pp. 431–440. doi: 10.1007/978-981-15-6876-3_33.

[23] B. M. Gayathri and C. P. Sumathi, “Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer,” in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Dec. 2016, pp. 1–5. doi: 10.1109/ICCIC.2016.7919576.

[24] M. Karabatak, “A new classifier for breast cancer detection based on Naïve Bayesian,” *Measurement*, vol. 72, pp. 32–36, Aug. 2015, doi: 10.1016/j.measurement.2015.04.028.

[25] M. F. Aslan, Y. Celik, K. Sabanci, and A. Durdu, “Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data,” *Int. J. Intell. Syst. Appl. Eng.*, vol. 6, no. 4, Art. no. 4, Dec. 2018, doi: 10.18201/ijisae.2018648455.

[26] S. Mojrian *et al.*, “Hybrid Machine Learning Model of Extreme Learning Machine Radial basis function for Breast Cancer Detection and Diagnosis; a Multilayer Fuzzy Expert System,” in *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, Oct. 2020, pp. 1–7. doi: 10.1109/RIVF48685.2020.9140744.

[27] R. Jeeva, S. Dhanasekar, A. Harshathunnisa, V. Eshwin, and A. Karn, “An Accurate Breast Cancer Detection and Classification using Image Processing,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 9, no. 3, p. 10, 2021, doi: 10.15680/IJIRCCE.2021.0903116.
[28] S. Ronoud and S. Asadi, “An evolutionary deep belief network extreme learning-based for breast cancer diagnosis,” *Soft Comput.*, vol. 23, no. 24, pp. 13139–13159, Dec. 2019, doi: 10.1007/s00500-019-03856-0.
[29] E. Aličković and A. Subasi, “Breast cancer diagnosis using GA feature selection and Rotation Forest,” *Neural Comput. Appl.*, vol. 28, no. 4, pp. 753–763, Apr. 2017, doi: 10.1007/s00521-015-2103-9.
[30] V. Chaurasia and S. Pal, “Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer,” *SN Comput. Sci.*, vol. 1, no. 5, p. 270, Aug. 2020, doi: 10.1007/s42979-020-00296-8.
[31] R. Soni, B, and S. Reddy, “Breast cancer detection by leveraging machine learning,” *ICT Express*, vol. 6. pp. 320–324, 2020.
[32] N. Liu, E.-S. Qi, M. Xu, B. Gao, and G.-Q. Liu, “A novel intelligent classification model for breast cancer diagnosis,” *Inf. Process. Manag.*, vol. 56, no. 3, pp. 609–623, May 2019, doi: 10.1016/j.ipm.2018.10.014.
[33] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, “Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets,” *J. Chin. Inst. Eng.*, vol. 43, no. 1, pp. 80–92, Jan. 2020, doi: 10.1080/02533839.2019.1676658.
[34] V. R. E. Christo, H. K. Nehemiah, J. Brighty, and A. Kannan, “Feature Selection and Instance Selection from Clinical Datasets Using Co-operative Co-evolution and Classification Using Random Forest,” *IETE J. Res.*, vol. 68, no. 4, pp. 2508–2521, Jul. 2022, doi: 10.1080/03772063.2020.1713917.
[35] M. F. Ak, “A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications,” *Healthcare*, vol. 8, no. 2, Art. no. 2, Jun. 2020, doi: 10.3390/healthcare8020111.
[36] H. Rao *et al.*, “Feature selection based on artificial bee colony and gradient boosting decision tree,” *Appl. Soft Comput.*, vol. 74, pp. 634–642, Jan. 2019, doi: 10.1016/j.asoc.2018.10.036.
[37] V. J. Kadam, S. M. Jadhav, and K. Vijayakumar, “Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Autoencoders and Softmax Regression,” *J. Med. Syst.*, vol. 43, no. 8, p. 263, Jul. 2019, doi: 10.1007/s10916-019-1397-z.
[38] M. Abdar *et al.*, “A new nested ensemble technique for automated diagnosis of breast cancer,” *Pattern Recognit. Lett.*, vol. 132, pp. 123–131, Apr. 2020, doi: 10.1016/j.patrec.2018.11.004.
[39] B. Sahu, S. Mohanty, and S. Rout, “A Hybrid Approach for Breast Cancer Classification and Diagnosis,” *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 6, no. 20, Jan. 2019, Accessed: Nov. 30, 2022. [Online]. Available: https://eudl.eu/doi/10.4108/eai.19-12-2018.156086
[40] R. A. Ibrahim, A. A. Ewees, D. Oliva, M. Abd Elaziz, and S. Lu, “Improved salp swarm algorithm based on particle swarm optimization for feature selection,” *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 8, pp. 3155–3169, Aug. 2019, doi: 10.1007/s12652-018-1031-9.
[41] E.-S. El-Kenawy and M. Eid, “Hybrid Gray Wolf and Particle Swarm Optimization for Feature Selection.” ICIC International 学会, 2020. doi: 10.24507/ijicic.16.03.831.

[42] S. J. Mambou, P. Maresova, O. Krejcar, A. Selamat, and K. Kuca, “Breast Cancer Detection Using Infrared Thermal Imaging and a Deep Learning Model,” *Sensors*, vol. 18, no. 9, Art. no. 9, Sep. 2018, doi: 10.3390/s18092799.
[43] S. Raiesdana, “Breast Cancer Detection Using Optimization-Based Feature Pruning and Classification Algorithms,” *Middle East J. Cancer*, vol. 12, no. 1, pp. 48–68, Jan. 2021, doi: 10.30476/mejc.2020.85601.1294.
[44] A. Ul Haq, J. Li, M. H. Memon, J. Khan, and S. Ud Din, “A novel integrated diagnosis method for breast cancer detection,” *J. Intell. Fuzzy Syst.*, vol. 38, no. 2, pp. 2383–2398, Jan. 2020, doi: 10.3233/JIFS-191461.
[45] G. Hamed, M. A. E.-R. Marey, S. E.-S. Amin, and M. F. Tolba, “Deep Learning in Breast Cancer Detection and Classification,” in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, A.-E. Hassanien, A. T. Azar, T. Gaber, D. Oliva, and F. M. Tolba, Eds., in Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020, pp. 322–333. doi: 10.1007/978-3-030-44289-7_30.
[46] S. B. Sakri, N. B. Abdul Rashid, and Z. Muhammad Zain, “Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction,” *IEEE Access*, vol. 6, pp. 29637–29647, 2018, doi: 10.1109/ACCESS.2018.2843443.
[47] M. A. Rahman, R. chandren Muniyandi, D. Albashish, M. M. Rahman, and O. L. Usman, “Artificial neural network with Taguchi method for robust classification model to improve classification accuracy of breast cancer,” *PeerJ Comput. Sci.*, vol. 7, p. e344, Jan. 2021, doi: 10.7717/peerj-cs.344.
[48] S. Wang, Y. Dai, J. Shen, and J. Xuan, “Research on expansion and classification of imbalanced data based on SMOTE algorithm,” *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Dec. 2021, doi: 10.1038/s41598-021-03430-5.
[49] G. Sahebi, P. Movahedi, M. Ebrahimi, T. Pahikkala, J. Plosila, and H. Tenhunen, “GeFeS: A generalized wrapper feature selection approach for optimizing classification performance,” *Comput. Biol. Med.*, vol. 125, p. 103974, Oct. 2020, doi: 10.1016/j.compbiomed.2020.103974.
[50] S. S. Tilwankar and B. S. Kirar, “Breast Cancer Detection using Principal Component Analysis and Machine Learning Models,” in *2021 First International Conference on Advances in Computing and Future Communication Technologies (ICACFCT)*, Dec. 2021, pp. 80–84. doi: 10.1109/ICACFCT53978.2021.9837342.
[51] Md. M. Islam, Md. R. Haque, H. Iqbal, Md. M. Hasan, M. Hasan, and M. N. Kabir, “Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques,” *SN Comput. Sci.*, vol. 1, no. 5, p. 290, Sep. 2020, doi: 10.1007/s42979-020-00305-w.
[52] M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath, and M. A. Bamiah, “Early detection of breast cancer using machine learning techniques,” *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no. 3–2. pp. 21–27, 2018.
[53] C. Boeri *et al.*, “Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation,” *Cancer Med.*, vol. 9, no. 9, pp. 3234–3243, 2020, doi: 10.1002/cam4.2811.
[54] N. Fatima, L. Liu, S. Hong, and H. Ahmed, “Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis,” *IEEE Access*, vol. 8, pp. 150360–150376, 2020, doi: 10.1109/ACCESS.2020.3016715.
[55] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, “Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review,” *Expert Syst. Appl.*, vol. 167, p. 114161, Apr. 2021, doi: 10.1016/j.eswa.2020.114161.

[56] Z. Khandezamin, M. Naderan, and M. J. Rashti, “Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier,” *J. Biomed. Inform.*, vol. 111, p. 103591, Nov. 2020, doi: 10.1016/j.jbi.2020.103591.
[57] F. Jiang, Q. Zhu, and T. Tian, “Breast Cancer Detection Based on Modified Harris Hawks Optimization and Extreme Learning Machine Embedded with Feature Weighting,” *Neural Process. Lett.*, Jan. 2022, doi: 10.1007/s11063-021-10700-w.
[58] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, “A support vector machine-based ensemble algorithm for breast cancer diagnosis,” *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, Jun. 2018, doi: 10.1016/j.ejor.2017.12.001.
[59] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, “An improved random forest-based rule extraction method for breast cancer diagnosis,” *Appl. Soft Comput.*, vol. 86, p. 105941, Jan. 2020, doi: 10.1016/j.asoc.2019.105941.
[60] P. K. P, M. A. B. V, and G. G. Nair, “An efficient classification framework for breast cancer using hyper parameter tuned Random Decision Forest Classifier and Bayesian Optimization,” *Biomed. Signal Process. Control*, vol. 68, p. 102682, Jul. 2021, doi: 10.1016/j.bspc.2021.102682.
[61] A. K. Naik, V. Kuppili, and D. R. Edla, “Efficient feature selection using one-pass generalized classifier neural network and binary bat algorithm with a novel fitness function,” *Soft Comput.*, vol. 24, no. 6, pp. 4575–4587, Mar. 2020, doi: 10.1007/s00500-019-04218-6.
[62] S. Dalwinder, S. Birmohan, and K. Manpreet, “Simultaneous feature weighting and parameter determination of Neural Networks using Ant Lion Optimization for the classification of breast cancer,” *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 337–351, Jan. 2020, doi: 10.1016/j.bbe.2019.12.004.
[63] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, “A knowledge-based system for breast cancer classification using fuzzy logic method,” *Telemat. Inform.*, vol. 34, no. 4, pp. 133–144, Jul. 2017, doi: 10.1016/j.tele.2017.01.007.
[64] M. H. Alshayeji, H. Ellethy, S. Abed, and R. Gupta, “Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach,” *Biomed. Signal Process. Control*, vol. 71, p. 103141, Jan. 2022, doi: 10.1016/j.bspc.2021.103141.
[65] M. Abdel-Basset, D. El-Shahat, I. El-henawy, V. H. C. de Albuquerque, and S. Mirjalili, “A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection,” *Expert Syst. Appl.*, vol. 139, p. 112824, Jan. 2020, doi: 10.1016/j.eswa.2019.112824.
[66] S. Jeyasingh and M. Veluchamy, “Modified Bat Algorithm for Feature Selection with the Wisconsin Diagnosis Breast Cancer (WDBC) Dataset,” *Asian Pac. J. Cancer Prev. APJCP*, vol. 18, no. 5, pp. 1257–1264, 2017, doi: 10.22034/APJCP.2017.18.5.1257.
[67] Nurhayati, F. Agustian, and M. D. I. Lubis, “Particle Swarm Optimization Feature Selection for Breast Cancer Prediction,” in *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, Oct. 2020, pp. 1–6. doi: 10.1109/CITSM50537.2020.9268865.
[68] J. Dheeba, N. Albert Singh, and S. Tamil Selvi, “Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach,” *J. Biomed. Inform.*, vol. 49, pp. 45–52, Jun. 2014, doi: 10.1016/j.jbi.2014.01.010.
[69] K. Rani, D. G.Naga Rama Devi, and L. Doddipalli, “Importance of Feature Extraction for Classification of Breast Cancer Datasets – A Study,” *Int. J. Sci. Innov. Math. Res.*, vol. 3, pp. 763–768, Jul. 2015.

[70] S. Murugesan, R. S. Bhuvaneswaran, H. Khanna Nehemiah, S. Keerthana Sankari, and Y. Nancy Jane, “Feature Selection and Classification of Clinical Datasets Using Bioinspired Algorithms and Super Learner,” *Comput. Math. Methods Med.*, vol. 2021, p. e6662420, May 2021, doi: 10.1155/2021/6662420.
[71] A. R. Vaka, B. Soni, and S. R. K., “Breast cancer detection by leveraging Machine Learning,” *ICT Express*, vol. 6, no. 4, pp. 320–324, Dec. 2020, doi: 10.1016/j.icte.2020.04.009.
[72] M. A. Rufai, A. S. Muhammad, S. Garba, and L. Audu, “MACHINE LEARNING MODEL FOR BREAST CANCER DETECTION,” *FUDMA J. Sci.*, vol. 4, no. 1, Art. no. 1, Apr. 2020.
[73] M. M. Saritas and A. B. Yaşar, “Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification,” *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, pp. 88–91, 2019.
[74] E. A. Bayrak, P. Kırcı, and T. Ensari, “Comparison of Machine Learning Methods for Breast Cancer Diagnosis,” in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, Apr. 2019, pp. 1–3. doi: 10.1109/EBBT.2019.8741990.
[75] S. Sharma, A. Aggarwal, and T. Choudhury, “Breast Cancer Detection Using Machine Learning Algorithms,” in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Dec. 2018, pp. 114–118. doi: 10.1109/CTEMS.2018.8769187.
[76] V. Bevilacqua, A. Brunetti, M. Triggiani, D. Magaletti, M. Telegrafo, and M. Moschetta, “An Optimized Feed-forward Artificial Neural Network Topology to Support Radiologists in Breast Lesions Classification,” in *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*, in GECCO ’16 Companion. New York, NY, USA: Association for Computing Machinery, Jul. 2016, pp. 1385–1392. doi: 10.1145/2908961.2931733.
[77] T. M. Mejía, M. G. Pérez, V. H. Andaluz, and A. Conci, “Automatic Segmentation and Analysis of Thermograms Using Texture Descriptors for Breast Cancer Detection,” in *2015 Asia-Pacific Conference on Computer Aided System Engineering*, Jul. 2015, pp. 24–29. doi: 10.1109/APCASE.2015.12.
[78] S. Aminikhanghahi, S. Shin, W. Wang, S. I. Jeon, S. H. Son, and C. Pack, “Study of wireless mammography image transmission impacts on robust cyber-aided diagnosis systems,” in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, in SAC ’15. New York, NY, USA: Association for Computing Machinery, Apr. 2015, pp. 2252–2256. doi: 10.1145/2695664.2695832.
[79] R. Sumbaly, “Diagnosis of Breast Cancer using Decision Tree Data Mining Technique,” *Int. J. Comput. Appl.*, vol. 98, p. 9.
[80] N. F. Idris and M. A. Ismail, “Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition,” *PeerJ Comput. Sci.*, vol. 7, p. e427, May 2021, doi: 10.7717/peerj-cs.427.
[81] H. Rajaguru and S. C. S R, “Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer,” *Asian Pac. J. Cancer Prev.*, vol. 20, no. 12, pp. 3777–3781, Dec. 2019, doi: 10.31557/APJCP.2019.20.12.3777.
[82] P. Lakhani and B. Sundaram, “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks,” *Radiology*, vol. 284, no. 2, pp. 574–582, Aug. 2017, doi: 10.1148/radiol.2017162326.
[83] K. Yasaka, H. Akai, O. Abe, and S. Kiryu, “Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A

Preliminary Study," *Radiology*, vol. 286, no. 3, pp. 887–896, Mar. 2018, doi: 10.1148/radiol.2017170706.
[84] P. F. Christ *et al.*, "Automatic Liver and Lesion Segmentation in CT Using Cascaded Fully Convolutional Neural Networks and 3D Conditional Random Fields," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 415–423. doi: 10.1007/978-3-319-46723-8_48.
[85] C. Krittanawong *et al.*, "Deep learning for cardiovascular medicine: a practical primer," *Eur. Heart J.*, vol. 40, no. 25, pp. 2058–2073, Jul. 2019, doi: 10.1093/eurheartj/ehz056.
[86] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Artificial Intelligence in Precision Cardiovascular Medicine," *J. Am. Coll. Cardiol.*, vol. 69, no. 21, pp. 2657–2664, May 2017, doi: 10.1016/j.jacc.2017.03.571.
[87] C. Krittanawong, A. S. Bomback, U. Baber, S. Bangalore, F. H. Messerli, and W. H. Wilson Tang, "Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension," *Curr. Hypertens. Rep.*, vol. 20, no. 9, p. 75, Jul. 2018, doi: 10.1007/s11906-018-0875-x.
[88] N. Nissa, S. Jamwal, and S. Ganie, "Heart Disease Prediction using Machine Learning Techniques," *Wesley. J. Res.*, vol. 13, no. 67, Mar. 2021.
[89] R. Alzubi, N. Ramzan, H. Alzoubi, and S. Katsigiannis, "SNPs-based Hypertension Disease Detection via Machine Learning Techniques," in *2018 24th International Conference on Automation and Computing (ICAC)*, Sep. 2018, pp. 1–6. doi: 10.23919/IConAC.2018.8748972.
[90] A. Dhankhar and S. Jain, "Prediction of Disease Using Machine Learning Algorithms," in *Smart and Sustainable Intelligent Systems*, John Wiley & Sons, Ltd, 2021, pp. 115–125. doi: 10.1002/9781119752134.ch8.
[91] P. Ghosh *et al.*, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
[92] E. Maini, B. Venkateswarlu, B. Maini, and D. Marwaha, "Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India," *Med. J. Armed Forces India*, vol. 77, no. 3, pp. 302–311, Jul. 2021, doi: 10.1016/j.mjafi.2020.10.013.
[93] S. Mishra, S. V. Neurkar, R. Patil, and S. Petkar, "Heart Disease Prediction System," *Int. J. Eng. Appl. Phys.*, vol. 1, no. 2, pp. 179–185, May 2021.
[94] B. Padmaja, C. Srinidhi, K. Sindhu, K. Vanaja, N. M. Deepika, and E. K. R. Patro, "Early and Accurate Prediction of Heart Disease Using Machine Learning Model," *Turk. J. Comput. Math. Educ. TURCOMAT*, vol. 12, no. 6, Art. no. 6, Jun. 2021.
[95] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Jan. 2021, pp. 1329–1333. doi: 10.1109/ICICT50816.2021.9358597.
[96] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Feb. 2020, pp. 1–5. doi: 10.1109/ic-ETITE47903.2020.242.
[97] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, and P. Ghuli, "Heart disease prediction using machine learning," *International Journal of Research and Technology*, vol. 9, no. 4. pp. 659–662, 2020.

[98] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
[99] D. Annepu and G. Gowtham, “Cardiovascular disease prediction using machine learning techniques,” *Int. Res. J. Eng. Technol.*, vol. 6, no. 4. pp. 3963–3971, 2019.
[100] H. B. F. David and S. A. Belcy, “HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES,” *ICTACT J. SOFT Comput.*, vol. 09, no. 01, p. 7, 2018, doi: 10.21917/ijsc.2018.0253.
[101] S. Nandhini, M. Debnath, A. Sharma, and Pushkar, “Heart disease prediction using machine learning,” *International Journal of Recent Engineering Research and Development (IJRERD)*, vol. 3, no. 10. pp. 39–46, 2018.
[102] M. N. R. Chowdhury, E. Ahmed, Md. A. D. Siddik, and A. U. Zaman, “Heart Disease Prognosis Using Machine Learning Classification Techniques,” in *2021 6th International Conference for Convergence in Technology (I2CT)*, Apr. 2021, pp. 1–6. doi: 10.1109/I2CT51068.2021.9418181.
[103] S. ware, S. K. Rakesh, and B. Choudhary, “Heart Attack Prediction by using Machine Learning Techniques,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 5, pp. 1577–1580, 2020, doi: 10.35940/ijrte.D9439.018520.
[104] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, “Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare,” *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
[105] R. M. Rishabh Magar and S. Raut, “Heart Disease Prediction Using Machine Learning,” *Journal of Emerging Technologies and Innovative Research*, vol. 7, no. 6. pp. 2081–2085, 2020.
[106] S. N. Khan *et al.*, “Comparative Analysis for Heart Disease Prediction,” *JOIV Int. J. Inform. Vis.*, vol. 1, no. 4–2, Art. no. 4–2, Nov. 2017, doi: 10.30630/joiv.1.4-2.66.
[107] A. Lakshmanarao, Y. Swathi, and S. Pullela, “Machine Learning Techniques For Heart Disease Prediction,” *Int. J. Sci. Technol. Res.*, vol. 8, p. 374, Oct. 2020.
[108] K. Hariharan, W. S. Vigneshwar, N. Sivaramakrishnan, and V. Subramaniyaswamy, “A COMPARATIVE STUDY ON HEART DISEASE ANALYSIS USING CLASSIFICATION TECHNIQUES,” *Int. J. Pure Appl. Math.*, vol. 119, no. 12e, pp. 13357–13366, 2018.
[109] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, “Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System,” *World J. Eng. Technol.*, vol. 6, no. 4, Art. no. 4, Sep. 2018, doi: 10.4236/wjet.2018.64057.
[110] H. Ayatollahi, L. Gholamhosseini, and M. Salehi, “Predicting coronary artery disease: a comparison between two data mining algorithms,” *BMC Public Health*, vol. 19, no. 1, p. 448, Apr. 2019, doi: 10.1186/s12889-019-6721-5.
[111] S. Kompella and V. Boddu, “Neural network based intelligent system for predicting heart disease,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, pp. 484–487, Jan. 2019.
[112] H. Meshref, “Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 12, 2019, doi: 10.14569/IJACSA.2019.0101236.
[113] C. N and M. B, “Classification Models on Cardiovascular Disease Prediction using Data Mining Techniques,” *Cardiovasc. Dis. Diagn.*, vol. 6, no. 6, pp. 1–4, 2018, doi: 10.4172/2329-9517.1000348.
[114] A. Sabay, L. Harris, V. Bejugama, and K. Jaceldo-Siegl, “Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data,” *SMU Data Sci. Rev.*,

vol. 1, no. 3, Aug. 2018, [Online]. Available: https://scholar.smu.edu/datasciencereview/vol1/iss3/12

[115] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Comput. Intell. Neurosci.*, vol. 2021, p. e8387680, Jul. 2021, doi: 10.1155/2021/8387680.

[116] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 6, p. 345, Oct. 2020, doi: 10.1007/s42979-020-00365-y.

[117] H. Arghandabi, "A Comparative Study of Machine Learning Algorithms for the Prediction of Heart Disease," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, no. 12, pp. 677–683, Dec. 2020, doi: 10.22214/ijraset.2020.32591.

[118] K. M. Almustafa, "Prediction of heart disease and classifiers' sensitivity analysis," *BMC Bioinformatics*, vol. 21, no. 1, p. 278, Jul. 2020, doi: 10.1186/s12859-020-03626-y.

[119] F. Rabbi *et al.*, "Performance Evaluation of Data Mining Classification Techniques for Heart Disease Prediction," *Am. J. Eng. Res.*, vol. 7, no. 2, pp. 278–283, 2018.

[120] S. 1 Geetha, C. P. 1 Devi, V. 1 Kalaivani, C. J. 1 Haritha, and G. 1 1 D. of I. T. Preetha, "Prediction Techniques of Heart Disease and Diabetes Disease using Machine Learning," *Turk. J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 3316–3325, 2021.

[121] X.-Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method," *Complexity*, vol. 2021, p. e6663455, Feb. 2021, doi: 10.1155/2021/6663455.

[122] A. Agrahary, "Heart Disease Prediction Using Machine Learning Algorithms," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 6, no. 4, pp. 137–149, Jul. 2020, doi: 10.32628/CSEIT206421.

[123] F. S. Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 10, no. 6, Art. no. 6, 29 2019, doi: 10.14569/IJACSA.2019.0100637.

[124] M. Tarawneh and O. Embarak, "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques," in *Advances in Internet, Data and Web Technologies*, L. Barolli, F. Xhafa, Z. A. Khan, and H. Odhabi, Eds., in Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing, 2019, pp. 447–454. doi: 10.1007/978-3-030-12839-5_41.

[125] S. Dhar, K. Roy, T. Dey, P. Datta, and A. Biswas, "A Hybrid Machine Learning Approach for Prediction of Heart Diseases," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Dec. 2018, pp. 1–6. doi: 10.1109/CCAA.2018.8777531.

[126] R. Prasad, P. Anjali, S. Adil, and N. Deepa, "Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 3, pp. 659–662.

[127] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," *Mob. Inf. Syst.*, vol. 2018, p. e3860146, Dec. 2018, doi: 10.1155/2018/3860146.

[128] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," in *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, Feb. 2020, pp. 452–457. doi: 10.1109/ICE348803.2020.9122958.

[129] H. Yao *et al.*, "Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests," *Front. Cell Dev. Biol.*, vol. 8, 2020, doi: doi.org/10.3389/fcell.2020.00683.
[130] L. N. Mahdy, K. A. Ezzat, H. H. Elmousalami, H. A. Ella, and A. E. Hassanien, "Automatic X-ray COVID-19 Lung Image Classification System based on Multi-Level Thresholding and Support Vector Machine." medRxiv, p. 2020.03.30.20047787, Apr. 06, 2020. doi: 10.1101/2020.03.30.20047787.
[131] P. K. Sethy and S. K. Behera, "Detection of Coronavirus Disease (COVID-19) Based on Deep Features." Preprints, Mar. 19, 2020. doi: 10.20944/preprints202003.0300.v1.
[132] L. Sun *et al.*, "Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19," *J. Clin. Virol.*, vol. 128, p. 104431, Jul. 2020, doi: 10.1016/j.jcv.2020.104431.
[133] M. Singh, S. Bansal, S. Ahuja, R. K. Dubey, B. K. Panigrahi, and N. Dey, "Transfer learning–based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data," *Med. Biol. Eng. Comput.*, vol. 59, no. 4, pp. 825–839, Apr. 2021, doi: 10.1007/s11517-020-02299-2.
[134] S. Singh, K. S. Parmar, S. J. S. Makkhan, J. Kaur, S. Peshoria, and J. Kumar, "Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries," *Chaos Solitons Fractals*, vol. 139, p. 110086, Oct. 2020, doi: 10.1016/j.chaos.2020.110086.
[135] M. Nour, Z. Cömert, and K. Polat, "A Novel Medical Diagnosis model for COVID-19 infection detection based on Deep Features and Bayesian Optimization," *Appl. Soft Comput.*, vol. 97, p. 106580, Dec. 2020, doi: 10.1016/j.asoc.2020.106580.
[136] H. Tabrizchi, A. Mosavi, A. Szabo-Gali, I. Felde, and L. Nadai, "Rapid COVID-19 Diagnosis Using Deep Learning of the Computerized Tomography Scans," in *2020 IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE)*, Nov. 2020, pp. 000173–000178. doi: 10.1109/CANDO-EPE51100.2020.9337794.
[137] H. Yue *et al.*, "Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study," *Ann. Transl. Med.*, vol. 8, no. 14, Art. no. 14, Jul. 2020, doi: 10.21037/atm-20-3026.
[138] W. Shi *et al.*, "A deep learning-based quantitative computed tomography model for predicting the severity of COVID-19: a retrospective study of 196 patients," *Ann. Transl. Med.*, vol. 9, no. 3, Art. no. 3, Feb. 2021, doi: 10.21037/atm-20-2464.
[139] L. Yan *et al.*, "An interpretable mortality prediction model for COVID-19 patients," *Nat. Mach. Intell.*, vol. 2, no. 5, Art. no. 5, May 2020, doi: 10.1038/s42256-020-0180-7.
[140] A. Salama, A. Darwsih, and A. E. Hassanien, "Artificial Intelligence Approach to Predict the COVID-19 Patient's Recovery," in *Digital Transformation and Emerging Technologies for Fighting COVID-19 Pandemic: Innovative Approaches*, A. E. Hassanien and A. Darwish, Eds., in Studies in Systems, Decision and Control. Cham: Springer International Publishing, 2021, pp. 121–133. doi: 10.1007/978-3-030-63307-3_8.
[141] R. Gupta, G. Pandey, P. Chaudhary, and S. Pal, "SEIR and Regression Model based COVID-19 outbreak predictions in India." medRxiv, p. 2020.04.01.20049825, Apr. 03, 2020. doi: 10.1101/2020.04.01.20049825.
[142] C. Hu *et al.*, "Early prediction of mortality risk among patients with severe COVID-19, using machine learning," *Int. J. Epidemiol.*, vol. 49, no. 6, pp. 1918–1929, Dec. 2020, doi: 10.1093/ije/dyaa171.

[143] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, and L. dos S. Coelho, “Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil,” *Chaos Solitons Fractals*, vol. 135, p. 109853, Jun. 2020, doi: 10.1016/j.chaos.2020.109853.
[144] M. Yadav, M. Perumal, and M. Srinivas, “Analysis on novel coronavirus (COVID-19) using machine learning methods,” *Chaos Solitons Fractals*, vol. 139, p. 110050, Oct. 2020, doi: 10.1016/j.chaos.2020.110050.
[145] J. Matos *et al.*, “Evaluation of novel coronavirus disease (COVID-19) using quantitative lung CT and clinical data: prediction of short-term outcome,” *Eur. Radiol. Exp.*, vol. 4, no. 1, p. 39, Jun. 2020, doi: 10.1186/s41747-020-00167-0.
[146] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, “Machine learning based approaches for detecting COVID-19 using clinical text data,” *Int. J. Inf. Technol.*, vol. 12, no. 3, pp. 731–739, Sep. 2020, doi: 10.1007/s41870-020-00495-9.
[147] H. S. Yang *et al.*, “Routine Laboratory Blood Tests Predict SARS-CoV-2 Infection Using Machine Learning,” *Clin. Chem.*, vol. 66, no. 11, pp. 1396–1404, Nov. 2020, doi: 10.1093/clinchem/hvaa200.
[148] M. Saqib, “Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model,” *Appl. Intell.*, vol. 51, no. 5, pp. 2703–2713, May 2021, doi: 10.1007/s10489-020-01942-7.
[149] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, “Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges,” *Int. J. Antimicrob. Agents*, vol. 55, no. 3, p. 105924, Mar. 2020, doi: 10.1016/j.ijantimicag.2020.105924.
[150] C. Iwendi *et al.*, “COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm,” *Front. Public Health*, vol. 8, 2020, doi: doi.org/10.3389/fpubh.2020.00357.
[151] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, “Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study,” *J. Med. Syst.*, vol. 44, no. 8, p. 135, Jul. 2020, doi: 10.1007/s10916-020-01597-4.
[152] T. Chakraborty and I. Ghosh, “Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis,” *Chaos Solitons Fractals*, vol. 135, p. 109850, Jun. 2020, doi: 10.1016/j.chaos.2020.109850.
[153] T. Tuncer, S. Dogan, and F. Ozyurt, “An automated Residual Exemplar Local Binary Pattern and iterative ReliefF based COVID-19 detection method using chest X-ray image,” *Chemom. Intell. Lab. Syst.*, vol. 203, p. 104054, Aug. 2020, doi: 10.1016/j.chemolab.2020.104054.
[154] R. M. Pereira, D. Bertolini, L. O. Teixeira, C. N. Silla, and Y. M. G. Costa, “COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios,” *Comput. Methods Programs Biomed.*, vol. 194, p. 105532, Oct. 2020, doi: 10.1016/j.cmpb.2020.105532.
[155] O. S. Albahri *et al.*, “Helping doctors hasten COVID-19 treatment: Towards a rescue framework for the transfusion of best convalescent plasma to the most critical patients based on biological requirements via ml and novel MCDM methods,” *Comput. Methods Programs Biomed.*, vol. 196, p. 105617, Nov. 2020, doi: 10.1016/j.cmpb.2020.105617.
[156] P. Wang, X. Zheng, J. Li, and B. Zhu, “Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics,” *Chaos Solitons Fractals*, vol. 139, p. 110058, Oct. 2020, doi: 10.1016/j.chaos.2020.110058.
[157] A. Abbasian Ardakani, U. R. Acharya, S. Habibollahi, and A. Mohammadi, “COVIDiag: a clinical CAD system to diagnose COVID-19 pneumonia based on CT

findings," *Eur. Radiol.*, vol. 31, no. 1, pp. 121–130, Jan. 2021, doi: 10.1007/s00330-020-07087-y.
[158] S. Wang *et al.*, "A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19)," *Eur. Radiol.*, vol. 31, no. 8, pp. 6096–6104, Aug. 2021, doi: 10.1007/s00330-021-07715-1.
[159] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," *Pattern Anal. Appl.*, vol. 24, no. 3, pp. 1207–1220, Aug. 2021, doi: 10.1007/s10044-021-00984-y.
[160] H. S. Maghdid, A. T. Asaad, K. Z. Ghafoor, A. S. Sadiq, S. Mirjalili, and M. K. Khan, "Diagnosing COVID-19 pneumonia from x-ray and CT images using deep learning and transfer learning algorithms," in *Multimodal Image Exploitation and Learning 2021*, SPIE, Apr. 2021, pp. 99–110. doi: 10.1117/12.2588672.
[161] B. Wang *et al.*, "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system," *Appl. Soft Comput.*, vol. 98, p. 106897, Jan. 2021, doi: 10.1016/j.asoc.2020.106897.
[162] J. Chen *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography," *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Nov. 2020, doi: 10.1038/s41598-020-76282-0.
[163] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, Jun. 2020, doi: 10.1007/s13246-020-00865-4.
[164] J. Zhang *et al.*, "Viral Pneumonia Screening on Chest X-ray Images Using Confidence-Aware Anomaly Detection." arXiv, Dec. 01, 2020. doi: 10.48550/arXiv.2003.12338.
[165] B. Ghoshal and A. Tucker, "Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection." arXiv, Mar. 27, 2020. doi: 10.48550/arXiv.2003.10769.
[166] S. Toraman, T. B. Alakus, and I. Turkoglu, "Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks," *Chaos Solitons Fractals*, vol. 140, p. 110122, Nov. 2020, doi: 10.1016/j.chaos.2020.110122.
[167] K. Hammoudi *et al.*, "Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19," *J. Med. Syst.*, vol. 45, no. 7, p. 75, Jun. 2021, doi: 10.1007/s10916-021-01745-4.
[168] M. Cifci, "Deep Learning Model for Diagnosis of Corona Virus Disease from CT Images," *Int. J. Sci. Res. Manag.*, vol. 11, no. 4, pp. 273–278, Apr. 2020.
[169] M. Loey, F. Smarandache, and N. E. M. Khalifa, "Within the Lack of Chest COVID-19 X-ray Dataset: A Novel Detection Model Based on GAN and Deep Transfer Learning," *Symmetry*, vol. 12, no. 4, Art. no. 4, Apr. 2020, doi: 10.3390/sym12040651.
[170] D. Singh, V. Kumar, null Vaishali, and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks," *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.*, vol. 39, no. 7, pp. 1379–1389, Jul. 2020, doi: 10.1007/s10096-020-03901-z.
[171] U. Özkaya, Ş. Öztürk, and M. Barstugan, "Coronavirus (COVID-19) Classification Using Deep Features Fusion and Ranking Technique," in *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach*, A.-E. Hassanien, N. Dey, and S. Elghamrawy, Eds., in Studies in Big Data. Cham: Springer International Publishing, 2020, pp. 281–295. doi: 10.1007/978-3-030-55258-9_17.

[172] M. Toğaçar, B. Ergen, and Z. Cömert, “COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches,” *Comput. Biol. Med.*, vol. 121, p. 103805, Jun. 2020, doi: 10.1016/j.compbiomed.2020.103805.

[173] Y. Pathak, P. K. Shukla, A. Tiwari, S. Stalin, S. Singh, and P. K. Shukla, “Deep Transfer Learning Based Classification Model for COVID-19 Disease,” *IRBM*, vol. 43, no. 2, pp. 87–92, Apr. 2022, doi: 10.1016/j.irbm.2020.05.003.

[174] N. Tsiknakis *et al.*, “Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays,” *Exp. Ther. Med.*, vol. 20, no. 2, pp. 727–735, Aug. 2020, doi: 10.3892/etm.2020.8797.

[175] M. E. H. Chowdhury *et al.*, “Can AI Help in Screening Viral and COVID-19 Pneumonia?,” *IEEE Access*, vol. 8, pp. 132665–132676, 2020, doi: 10.1109/ACCESS.2020.3010287.

[176] M. Rahimzadeh and A. Attar, “A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2,” *Inform. Med. Unlocked*, vol. 19, p. 100360, Jan. 2020, doi: 10.1016/j.imu.2020.100360.

[177] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi, “Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks,” *Comput. Biol. Med.*, vol. 121, p. 103795, Jun. 2020, doi: 10.1016/j.compbiomed.2020.103795.

[178] K. El Asnaoui, Y. Chawki, and A. Idri, “Automated Methods for Detection and Classification Pneumonia Based on X-Ray Images Using Deep Learning,” in *Artificial Intelligence and Blockchain for Future Cybersecurity Applications*, Y. Maleh, Y. Baddi, M. Alazab, L. Tawalbeh, and I. Romdhani, Eds., in Studies in Big Data. Cham: Springer International Publishing, 2021, pp. 257–284. doi: 10.1007/978-3-030-74575-2_14.

[179] S. Rajaraman and S. Antani, “Weakly Labeled Data Augmentation for Deep Learning: A Study on COVID-19 Detection in Chest X-Rays,” *Diagnostics*, vol. 10, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/diagnostics10060358.

[180] V. S. Lakshmi, V. Nithya, K. Sripriya, C. Preethi, and K. Logeshwari, “Prediction of Diabetes Patient Stage Using Ontology Based Machine Learning System,” in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, Mar. 2019, pp. 1–4. doi: 10.1109/ICSCAN.2019.8878831.

[181] H. R. Divakar, B. R. Prakash, and M. Mamatha, “An Ontology Based System for Healthcare People to Prevent Cardiovascular Diseases,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 2S11, pp. 983–988, Nov. 2019, doi: 10.35940/ijrte.B1164.0982S1119.

[182] S. Pc, V. Pv, R. Krishnan, and Y. Saad, “ONTOLOGY DRIVEN ANALYSIS AND PREDICTION OF PATIENT RISK IN DIABETES,” *Can. J. Pure Appl. Sci.*, vol. 8, no. 3, pp. 3043–3050, Oct. 2014.

[183] A. Verma, N. Kasabov, E. Rush, and Q. Song, “Ontology Based Personalized Modeling for Chronic Disease Risk Analysis: An Integrated Approach,” in *Advances in Neuro-Information Processing*, M. Köppen, N. Kasabov, and G. Coghill, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 1204–1210. doi: 10.1007/978-3-642-02490-0_146.

[184] P. C. Sherimon and R. Krishnan, “OntoDiabetic: An Ontology-Based Clinical Decision Support System for Diabetic Patients,” *Arab. J. Sci. Eng.*, vol. 41, no. 3, pp. 1145–1160, Mar. 2016, doi: 10.1007/s13369-015-1959-4.

[185] D. Parry and J. MacRae, “Fuzzy ontologies for cardiovascular risk prediction - A research approach,” in *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Jul. 2013, pp. 1–4. doi: 10.1109/FUZZ-IEEE.2013.6622564.

[186] S. Althubaiti *et al.*, “Ontology-based prediction of cancer driver genes,” *Sci. Rep.*, vol. 9, no. 1, Art. no. 1, Nov. 2019, doi: 10.1038/s41598-019-53454-1.
[187] S. N. Raju, K. Snehaja, and B. Srinivas, “Ontology Based Disease Prediction System,” in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, Jul. 2021, pp. 1–6. doi: 10.1109/ICCES51350.2021.9489132.
[188] H. Mahmoud, E. Abbas, and I. Fathy, “Data mining and ontology-based techniques in healthcare management,” *Int. J. Intell. Eng. Inform.*, vol. 6, no. 6, p. 509, 2018, doi: 10.1504/IJIEI.2018.096549.
[189] P. L. Chavan and P. M. S. Karyakarte, “Ontology Based System for Prediction of Diseases,” *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 7, no. 3, pp. 277–285, Jun. 2020, doi: 10.32628/IJSRSET207365.
[190] University of Babylon, M. H. Jabardi, and A. S. Hadi, “Ontology Meter for Twitter Fake Accounts Detection,” *Int. J. Intell. Eng. Syst.*, vol. 14, no. 1, pp. 410–419, Feb. 2021, doi: 10.22266/ijies2021.0228.38.
[191] Z. M. A. Khan, S. Saeidlou, and M. Saadat, “Ontology-based decision tree model for prediction in a manufacturing network,” *Prod. Manuf. Res.*, vol. 7, no. 1, pp. 335–349, Jan. 2019, doi: 10.1080/21693277.2019.1621228.
[192] M. JABARDI and A. Hadi, “Twitter Fake Account Detection and Classification using Ontological Engineering and Semantic Web Rule Language,” *Karbala Int. J. Mod. Sci.*, vol. 6, no. 4, Dec. 2020, doi: 10.33640/2405-609X.2285.
[193] Q. Cao, A. Samet, C. Zanni-Merk, F. de B. de Beuvron, and C. Reich, “An Ontology-based Approach for Failure Classification in Predictive Maintenance Using Fuzzy C-means and SWRL Rules,” *Procedia Comput. Sci.*, vol. 159, pp. 630–639, Jan. 2019, doi: 10.1016/j.procs.2019.09.218.
[194] K. Buchan, M. Filannino, and Ö. Uzuner, “Automatic prediction of coronary artery disease from clinical narratives,” *J. Biomed. Inform.*, vol. 72, pp. 23–32, Aug. 2017, doi: 10.1016/j.jbi.2017.06.019.
[195] S. Y. Banihashem and S. Shishehchi, “Ontology-Based decision tree model for prediction of fatty liver diseases,” *Comput. Methods Biomech. Biomed. Engin.*, vol. 0, no. 0, pp. 1–11, May 2022, doi: 10.1080/10255842.2022.2081502.
[196] C. Trevisan, G. Sergi, and S. Maggi, “Gender Differences in Brain-Heart Connection,” in *Brain and Heart Dynamics*, S. Govoni, P. Politi, and E. Vanoli, Eds., Cham: Springer International Publishing, 2020, pp. 937–951. doi: 10.1007/978-3-030-28008-6_61.
[197] I. Yekkala and S. Dixit, “Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection,” *Int. J. Big Data Anal. Healthc. IJBDAH*, vol. 3, no. 1, pp. 1–12, 2018, doi: 10.4018/IJBDAH.2018010101.
[198] “Sex, Age, Cardiovascular Risk Factors, and Coronary Heart Disease | Circulation.” https://www.ahajournals.org/doi/full/10.1161/01.cir.99.9.1165 (accessed Aug. 10, 2022).
[199] K. Uyar and A. İlhan, “Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks,” *Procedia Comput. Sci.*, vol. 120, pp. 588–593, Jan. 2017, doi: 10.1016/j.procs.2017.11.283.
[200] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, “A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease,” in *2017 IEEE Symposium on Computers and Communications (ISCC)*, Jul. 2017, pp. 204–207. doi: 10.1109/ISCC.2017.8024530.
[201] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, “Innovative Artificial Neural Networks-Based Decision Support System for Heart Diseases Diagnosis,” *J. Intell. Learn. Syst. Appl.*, vol. 5, no. 3, Art. no. 3, Jul. 2013, doi: 10.4236/jilsa.2013.53019.

[202] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *J. Appl. Biomed.*, vol. 11, no. 2, pp. 47–58, Jan. 2013, doi: 10.2478/v10136-012-0031-x.

[203] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telemat. Inform.*, vol. 36, pp. 82–93, Mar. 2019, doi: 10.1016/j.tele.2018.11.007.

[204] N. Kausar, S. Palaniappan, B. B. Samir, A. Abdullah, and N. Dey, "Systematic Analysis of Applied Data Mining Based Optimization Algorithms in Clinical Attribute Extraction and Classification for Diagnosis of Cardiac Patients," in *Applications of Intelligent Optimization in Biology and Medicine: Current Trends and Open Problems*, A.-E. Hassanien, C. Grosan, and M. Fahmy Tolba, Eds., in Intelligent Systems Reference Library. Cham: Springer International Publishing, 2016, pp. 217–231. doi: 10.1007/978-3-319-21212-8_9.

[205] Md. M. Alam *et al.*, "D-CARE: A Non-invasive Glucose Measuring Technique for Monitoring Diabetes Patients," in *Proceedings of International Joint Conference on Computational Intelligence*, M. S. Uddin and J. C. Bansal, Eds., in Algorithms for Intelligent Systems. Singapore: Springer, 2020, pp. 443–453. doi: 10.1007/978-981-13-7564-4_38.

[206] M. Ashraf *et al.*, "Prediction of Cardiovascular Disease Through Cutting-Edge Deep Learning Technologies: An Empirical Study Based on TENSORFLOW, PYTORCH and KERAS," in *International Conference on Innovative Computing and Communications*, D. Gupta, A. Khanna, S. Bhattacharyya, A. E. Hassanien, S. Anand, and A. Jaiswal, Eds., in Advances in Intelligent Systems and Computing. Singapore: Springer, 2021, pp. 239–255. doi: 10.1007/978-981-15-5113-0_18.

[207] F. Andreotti *et al.*, "Prediction of the onset of cardiovascular diseases from electronic health records using multi-task gated recurrent units." arXiv, Jul. 16, 2020. doi: 10.48550/arXiv.2007.08491.

[208] W. Wiharto, H. Kusnanto, and H. Herianto, "Hybrid System of Tiered Multivariate Analysis and Artificial Neural Network for Coronary Heart Disease Diagnosis," *Int. J. Electr. Comput. Eng. IJECE*, vol. 7, no. 2, Art. no. 2, Apr. 2017, doi: 10.11591/ijece.v7i2.pp1023-1031.

[209] A. K. Paul, P. C. Shill, Md. R. I. Rabin, and M. A. H. Akhand, "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease," in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, May 2016, pp. 145–150. doi: 10.1109/ICIEV.2016.7759984.

[210] X. Liu *et al.*, "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method," *Comput. Math. Methods Med.*, vol. 2017, p. e8272091, Jan. 2017, doi: 10.1155/2017/8272091.

[211] L. Yahaya, N. D. Oye, and E. J. Garba, "A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques," *Am. J. Artif. Intell.*, vol. 4, no. 1, Art. no. 1, Apr. 2020, doi: 10.11648/j.ajai.20200401.12.

[212] M. Shouman, T. Turner, and R. Stocker, "Integrating Clustering with Different Data Mining Techniques in the Diagnosis of Heart Disease," *J. Comput. Sci. Eng.*, vol. 20, no. 1, p. 11, 2013.

[213] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Inform. Med. Unlocked*, vol. 20, p. 100402, Jan. 2020, doi: 10.1016/j.imu.2020.100402.

[214] B. A. Tama, S. Im, and S. Lee, "Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble," *BioMed Res. Int.*, vol. 2020, p. e9816142, Apr. 2020, doi: 10.1155/2020/9816142.

[215] J. Mishra and S. Tarar, “Chronic Disease Prediction Using Deep Learning,” in *Advances in Computing and Data Sciences*, M. Singh, P. K. Gupta, V. Tyagi, J. Flusser, T. Ören, and G. Valentino, Eds., in Communications in Computer and Information Science. Singapore: Springer, 2020, pp. 201–211. doi: 10.1007/978-981-15-6634-9_19.

[216] X. Zenuni, B. Raufi, F. Ismaili, and J. Ajdari, “State of the Art of Semantic Web for Healthcare,” *Procedia - Soc. Behav. Sci.*, vol. 195, pp. 1990–1998, Jul. 2015, doi: 10.1016/j.sbspro.2015.06.213.

[217] F. B. Nardon and L. A. Moura, “Knowledge Sharing and Information Integration in Healthcare using Ontologies and Deductive Databases,” *MEDINFO 2004*, pp. 62–66, 2004, doi: 10.3233/978-1-60750-949-3-62.

[218] A. Kilic, “Artificial Intelligence and Machine Learning in Cardiovascular Health Care,” *Ann. Thorac. Surg.*, vol. 109, no. 5, pp. 1323–1329, May 2020, doi: 10.1016/j.athoracsur.2019.09.042.

[219] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, “Machine learning in cardiovascular medicine: are we there yet?,” *Heart*, vol. 104, no. 14, pp. 1156–1164, Jul. 2018, doi: 10.1136/heartjnl-2017-311198.

[220] F. Z. Abdeldjouad, M. Brahami, and N. Matta, “A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques,” in *The Impact of Digital Technologies on Public Health in Developed and Developing Countries*, M. Jmaiel, M. Mokhtari, B. Abdulrazak, H. Aloulou, and S. Kallel, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 299–306. doi: 10.1007/978-3-030-51517-1_26.

[221] H. E. Hamdaoui, S. Boujraf, N. E. H. Chaoui, B. Alami, and M. Maaroufi, “Improving Heart Disease Prediction Using Random Forest and AdaBoost Algorithms,” *Int. J. Online Biomed. Eng. IJOE*, vol. 17, no. 11, Art. no. 11, Nov. 2021, doi: 10.3991/ijoe.v17i11.24781.

[222] J. A. Quesada *et al.*, “Machine learning to predict cardiovascular risk,” *Int. J. Clin. Pract.*, vol. 73, no. 10, p. e13389, 2019, doi: 10.1111/ijcp.13389.

[223] C. Krittanawong *et al.*, “Machine learning prediction in cardiovascular diseases: a meta-analysis,” *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Sep. 2020, doi: 10.1038/s41598-020-72685-1.

[224] N. N. Anuar *et al.*, “Cardiovascular Disease Prediction from Electrocardiogram by Using Machine Learning,” *Int. J. Online Biomed. Eng. IJOE*, vol. 16, no. 07, Art. no. 07, Jun. 2020, doi: 10.3991/ijoe.v16i07.13569.

[225] S. M. H. S. Iqbal, N. Jahan, A. S. Moni, and M. Khatun, “An Effective Analytics and Performance Measurement of different Machine Learning Algorithms for Predicting Heart Disease,” *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 13, no. 2, Art. no. 2, 51/28 2022, doi: 10.14569/IJACSA.2022.0130250.

[226] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, “Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison,” *Comput. Biol. Med.*, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

[227] M. Swathy and K. Saruladha, “A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques,” *ICT Express*, Sep. 2021, doi: 10.1016/j.icte.2021.08.021.

[228] A. K. Faieq and M. M. Mijwil, “Prediction of of heart diseases utilising support vector machine and artificial neural network,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 1, Art. no. 1, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp374-380.

[229] D. Shah, S. Patel, and S. K. Bharti, “Heart Disease Prediction using Machine Learning Techniques,” *SN Comput. Sci.*, vol. 1, no. 6, p. 345, Oct. 2020, doi: 10.1007/s42979-020-00365-y.
[230] A. Alfaidi, R. Aljuhani, B. Alshehri, H. Alwadei, and S. Sabbeh, “Machine Learning: Assisted Cardiovascular Diseases Diagnosis,” *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 13, no. 2, Art. no. 2, 51/28 2022, doi: 10.14569/IJACSA.2022.0130216.
[231] A. Kondababu, V. Siddhartha, BHK. B. Kumar, and B. Penumutchi, “A comparative study on machine learning based heart disease prediction,” *Mater. Today Proc.*, Feb. 2021, doi: 10.1016/j.matpr.2021.01.475.
[232] R. R. K. AL-Taie, B. J. Saleh, A. Y. F. Saedi, and L. A. Salman, “Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq,” *Int. J. Electr. Comput. Eng. IJECE*, vol. 11, no. 6, Art. no. 6, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5229-5239.
[233] B. J. Saleh, R. R. K. Al_Taie, and A. A. Mhawes, “Machine Learning Architecture for Heart Disease Detection: A Case Study in Iraq,” *Int. J. Online Biomed. Eng. IJOE*, vol. 18, no. 02, Art. no. 02, Feb. 2022, doi: 10.3991/ijoe.v18i02.27143.
[234] Aman and R. S. Chhillar, “Analyzing Predictive Algorithms in Data Mining for Cardiovascular Disease using WEKA Tool,” *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 12, no. 8, Art. no. 8, 31 2021, doi: 10.14569/IJACSA.2021.0120817.
[235] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, “An artificial intelligence model for heart disease detection using machine learning algorithms,” *Healthc. Anal.*, vol. 2, p. 100016, Nov. 2022, doi: 10.1016/j.health.2022.100016.
[236] H. Meshref, “Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach,” *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 10, no. 12, Art. no. 12, Jun. 2019, doi: 10.14569/IJACSA.2019.0101236.
[237] S. Krishnan, P. Magalingam, and R. Ibrahim, “Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction,” *Int. J. Electr. Comput. Eng. IJECE*, vol. 11, no. 6, Art. no. 6, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5467-5476.
[238] H. Bensenane, D. Aksa, F. W. Omari, and A. Rahmoun, “A deep learning-based cardio-vascular disease diagnosis system,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 2, Art. no. 2, Feb. 2022, doi: 10.11591/ijeecs.v25.i2.pp963-971.
[239] Z. Sabouri, Y. Maleh, and N. Gherabi, “Benchmarking Classification Algorithms for Measuring the Performance on Maintainable Applications,” in *Advances in Information, Communication and Cybersecurity*, Y. Maleh, M. Alazab, N. Gherabi, L. Tawalbeh, and A. A. Abd El-Latif, Eds., in Lecture Notes in Networks and Systems. Cham: Springer International Publishing, 2022, pp. 173–179. doi: 10.1007/978-3-030-91738-8_17.
[240] M. A. Musen, “The protégé project: a look back and a look forward,” *AI Matters*, vol. 1, no. 4, pp. 4–12, Jun. 2015, doi: 10.1145/2757001.2757003.
[241] L. Rubinger, A. Gazendam, S. Ekhtiari, and M. Bhandari, “Machine learning and artificial intelligence in research and healthcare,” *Injury*, Feb. 2022, doi: 10.1016/j.injury.2022.01.046.
[242] H. Abdel-Jaber, D. Devassy, A. Al Salam, L. Hidaytallah, and M. EL-Amir, “A Review of Deep Learning Algorithms and Their Applications in Healthcare,” *Algorithms*, vol. 15, no. 2, Art. no. 2, Feb. 2022, doi: 10.3390/a15020071.
[243] E. Dong *et al.*, “The Johns Hopkins University Center for Systems Science and Engineering COVID-19 Dashboard: data collection process, challenges faced, and lessons learned,” *Lancet Infect. Dis.*, Aug. 2022, doi: 10.1016/S1473-3099(22)00434-0.
[244] Z. Y. Zu *et al.*, “Coronavirus Disease 2019 (COVID-19): A Perspective from China,” *Radiology*, vol. 296, no. 2, pp. E15–E25, Aug. 2020, doi: 10.1148/radiol.2020200490.

[245] A. Bernheim *et al.*, “Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection,” *Radiology*, p. 200463, Feb. 2020, doi: 10.1148/radiol.2020200463.
[246] G. Saranya and A. Pravin, “A comprehensive study on disease risk predictions in machine learning,” *Int. J. Electr. Comput. Eng. IJECE*, vol. 10, no. 4, Art. no. 4, Aug. 2020, doi: 10.11591/ijece.v10i4.pp4217-4225.
[247] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and Machine Learning forecasting methods: Concerns and ways forward,” *PLOS ONE*, vol. 13, no. 3, p. e0194889, Mar. 2018, doi: 10.1371/journal.pone.0194889.
[248] G. Bontempi, S. Ben Taieb, and Y.-A. Le Borgne, “Machine Learning Strategies for Time Series Forecasting,” in *Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures*, M.-A. Aufaure and E. Zimányi, Eds., in Lecture Notes in Business Information Processing. Berlin, Heidelberg: Springer, 2013, pp. 62–77. doi: 10.1007/978-3-642-36318-4_3.
[249] F. J. Harrell, K. L. Lee, D. B. Matchar, and T. A. Reichert, “Regression models for prognostic prediction: advantages, problems, and suggested solutions,” *Cancer Treat. Rep.*, vol. 69, no. 10, pp. 1071–1077, Oct. 1985.
[250] P. Lapuerta, S. P. Azen, and L. Labree, “Use of Neural Networks in Predicting the Risk of Coronary Artery Disease,” *Comput. Biomed. Res.*, vol. 28, no. 1, pp. 38–52, Feb. 1995, doi: 10.1006/cbmr.1995.1004.
[251] K. M. Anderson, P. M. Odell, P. W. F. Wilson, and W. B. Kannel, “Cardiovascular disease risk profiles,” *Am. Heart J.*, vol. 121, no. 1, Part 2, pp. 293–298, Jan. 1991, doi: 10.1016/0002-8703(91)90861-B.
[252] H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, M. Bahaj, and M. R. Naqvi, “The Impact of Ontology on the Prediction of Cardiovascular Disease Compared to Machine Learning Algorithms,” *Int. J. Online Biomed. Eng. IJOE*, vol. 18, no. 11, Art. no. 11, Aug. 2022, doi: 10.3991/ijoe.v18i11.32647.
[253] H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, F. Qanouni, and M. Bahaj, “An Ontological Model based on Machine Learning for Predicting Breast Cancer,” *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 13, no. 7, Art. no. 7, 31 2022, doi: 10.14569/IJACSA.2022.0130715.
[254] F. Petropoulos and S. Makridakis, “Forecasting the novel coronavirus COVID-19,” *PLOS ONE*, vol. 15, no. 3, p. e0231236, Mar. 2020, doi: 10.1371/journal.pone.0231236.
[255] G. Grasselli, A. Pesenti, and M. Cecconi, “Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response,” *JAMA*, vol. 323, no. 16, pp. 1545–1546, Apr. 2020, doi: 10.1001/jama.2020.4031.
[256] L. Wynants *et al.*, “Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal,” *BMJ*, vol. 369, p. m1328, Apr. 2020, doi: 10.1136/bmj.m1328.
[257] Y. Xiang, Y. Jia, L. Chen, L. Guo, B. Shu, and E. Long, “COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models,” *Infect. Dis. Model.*, vol. 6, pp. 324–342, Jan. 2021, doi: 10.1016/j.idm.2021.01.001.
[258] Y. Meraihi, A. B. Gabis, S. Mirjalili, A. Ramdane-Cherif, and F. E. Alsaadi, “Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey,” *SN Comput. Sci.*, vol. 3, no. 4, p. 286, May 2022, doi: 10.1007/s42979-022-01184-z.
[259] L. W. Mary and S. A. A. Raj, “Machine Learning Algorithms for Predicting SARS-CoV-2 (COVID-19) – A Comparative Analysis,” in *2021 2nd International Conference*

on Smart Electronics and Communication (ICOSEC), Oct. 2021, pp. 1607–1611. doi: 10.1109/ICOSEC51865.2021.9591801.

[260] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," *Big Data Min. Anal.*, vol. 4, no. 2, pp. 116–123, Jun. 2021, doi: 10.26599/BDMA.2020.9020016.

[261] C. N. Villavicencio, J. J. E. Macrohon, X. A. Inbaraj, J.-H. Jeng, and J.-G. Hsieh, "COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA," *Algorithms*, vol. 14, no. 7, Art. no. 7, Jul. 2021, doi: 10.3390/a14070201.

[262] P. Guleria, S. Ahmed, A. Alhumam, and P. N. Srinivasu, "Empirical Study on Classifiers for Earlier Prediction of COVID-19 Infection Cure and Death Rate in the Indian States," *Healthcare*, vol. 10, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/healthcare10010085.

[263] R. Kumari *et al.*, "Analysis and predictions of spread, recovery, and death caused by COVID-19 in India," *Big Data Min. Anal.*, vol. 4, no. 2, pp. 65–75, Jun. 2021, doi: 10.26599/BDMA.2020.9020013.

[264] S. Bhardwaj, H. Bhardwaj, J. Bhardwaj, and P. Gupta, "Global Prediction of COVID-19 Cases and Deaths using Machine Learning," in *2021 Sixth International Conference on Image Information Processing (ICIIP)*, Nov. 2021, pp. 422–426. doi: 10.1109/ICIIP53038.2021.9702560.

[265] K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad, and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting COVID-19 mortality," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 2, Jan. 2022, doi: 10.1186/s12911-021-01742-0.

[266] K. B. Prakash, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 2199–2204, May 2020, doi: 10.30534/ijeter/2020/117852020.

[267] P. K. Roy and A. Kumar, "Early prediction of COVID-19 using ensemble of transfer learning," *Comput. Electr. Eng.*, vol. 101, p. 108018, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108018.

[268] J. K. Shade *et al.*, "Real-Time Prediction of Mortality, Cardiac Arrest, and Thromboembolic Complications in Hospitalized Patients With COVID-19," *JACC Adv.*, vol. 1, no. 2, p. 100043, Jun. 2022, doi: 10.1016/j.jacadv.2022.100043.

[269] S. A.-F. Sayed, A. M. Elkorany, and S. Sayed Mohammad, "Applying Different Machine Learning Techniques for Prediction of COVID-19 Severity," *IEEE Access*, vol. 9, pp. 135697–135707, 2021, doi: 10.1109/ACCESS.2021.3116067.

[270] S. Bhutia, B. Patra, and M. Ray, "COVID-19 epidemic: analysis and prediction," *IAES Int. J. Artif. Intell. IJ-AI*, vol. 11, no. 2, Art. no. 2, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp736-745.

[271] A. H. Ahmed, M. N. A. Al-Hamadani, and I. A. Satam, "Prediction of COVID-19 disease severity using machine learning techniques," *Bull. Electr. Eng. Inform.*, vol. 11, no. 2, Art. no. 2, Apr. 2022, doi: 10.11591/eei.v11i2.3272.

[272] N. Hayati, F. Fauziah, D. R. Poetra, and D. Wandi, "Trend of the spread of COVID-19 in Indonesia using the machine learning prophet algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 24, no. 3, Art. no. 3, Dec. 2021, doi: 10.11591/ijeecs.v24.i3.pp1780-1788.

[273] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *Am. Stat.*, vol. 46, no. 3, pp. 175–185, Aug. 1992, doi: 10.1080/00031305.1992.10475879.

[274] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Aug. 1995, pp. 278–282 vol.1. doi: 10.1109/ICDAR.1995.598994.
[275] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[276] A. Moubayed, M. Injadat, A. B. Nassif, H. Lutfiyya, and A. Shami, “E-Learning: Challenges and Research Opportunities Using Machine Learning & Data Analytics,” *IEEE Access*, vol. 6, pp. 39117–39138, 2018, doi: 10.1109/ACCESS.2018.2851790.
[277] M. Mishra and M. Srivastava, “A view of Artificial Neural Network,” in *2014 International Conference on Advances in Engineering & Technology Research (ICAETR - 2014)*, Aug. 2014, pp. 1–3. doi: 10.1109/ICAETR.2014.7012785.
[278] A. Jentzen and P. von Wurstemberger, “Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates,” *J. Complex.*, vol. 57, p. 101438, Apr. 2020, doi: 10.1016/j.jco.2019.101438.
[279] S. Sharma, J. Agrawal, S. Agarwal, and S. Sharma, “Machine learning techniques for data mining: A survey,” in *2013 IEEE International Conference on Computational Intelligence and Computing Research*, Dec. 2013, pp. 1–6. doi: 10.1109/ICCIC.2013.6724149.
[280] M. Kuhkan, “A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm,” *Int. J. Comput. Eng. Inf. Technol.*, vol. 8, no. 6, pp. 90–95, Jun. 2016.
[281] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, “Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification,” in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Nov. 2017, pp. 294–298. doi: 10.1109/ICITISEE.2017.8285514.
[282] A. A. Soofi and A. Awan, “Classification Techniques in Machine Learning: Applications and Issues,” *J. Basic Appl. Sci.*, vol. 13, pp. 459–465, Jan. 2017, doi: 10.6000/1927-5129.2017.13.76.
[283] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Clim. Res.*, vol. 30, no. 1, pp. 79–82, Dec. 2005, doi: 10.3354/cr030079.
[284] M. Mori *et al.*, “Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-field digital mammography in a population of women with dense breasts,” *Breast Cancer Tokyo Jpn.*, vol. 24, no. 1, pp. 104–110, Jan. 2017, doi: 10.1007/s12282-016-0681-8.
[285] N. Harbeck *et al.*, “Breast cancer,” *Nat. Rev. Dis. Primer*, vol. 5, no. 1, Art. no. 1, Sep. 2019, doi: 10.1038/s41572-019-0111-2.
[286] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015, doi: 10.1016/j.csbj.2014.11.005.
[287] M. Shi and B. Zhang, “Semi-supervised learning improves gene expression-based prediction of cancer recurrence,” *Bioinformatics*, vol. 27, no. 21, pp. 3017–3023, Nov. 2011, doi: 10.1093/bioinformatics/btr502.
[288] S. Becker, “A historic and scientific review of breast cancer: The next global healthcare challenge,” *Int. J. Gynecol. Obstet.*, vol. 131, no. S1, pp. S36–S39, 2015, doi: 10.1016/j.ijgo.2015.03.015.
[289] A. R. Padhani *et al.*, “Diffusion-Weighted Magnetic Resonance Imaging as a Cancer Biomarker: Consensus and Recommendations,” *Neoplasia*, vol. 11, no. 2, pp. 102–125, Feb. 2009, doi: 10.1593/neo.81328.

[290] D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, Jun. 2005, doi: 10.1016/j.artmed.2004.07.002.
[291] K. Rani, “Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique,” *Int. J. Comput. Appl.*, vol. 10, Nov. 2010, doi: 10.5120/1465-1980.
[292] A. S. Sarvestani, A. A. Safavi, N. M. Parandeh, and M. Salehi, “Predicting breast cancer survivability using data mining techniques,” in *2010 2nd International Conference on Software Technology and Engineering*, Oct. 2010, pp. V2-227-V2-231. doi: 10.1109/ICSTE.2010.5608818.
[293] C. Sotiriou *et al.*, “Breast cancer classification and prognosis based on gene expression profiles from a population-based study,” *Proc. Natl. Acad. Sci.*, vol. 100, no. 18, pp. 10393–10398, Sep. 2003, doi: 10.1073/pnas.1732912100.
[294] C. Shravya, K. Pravalika, and S. Subhani, “Prediction of Breast Cancer Using Supervised Machine Learning Techniques,” vol. 8, no. 6, p. 5, 2019.
[295] S. I. Ayon, Md. M. Islam, and Md. R. Hossain, “Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques,” *IETE J. Res.*, vol. 68, no. 4, pp. 2488–2507, Jul. 2022, doi: 10.1080/03772063.2020.1713916.
[296] L. J. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, “Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients’ Recovery,” *SN Comput. Sci.*, vol. 1, no. 4, p. 206, 2020, doi: 10.1007/s42979-020-00216-w.
[297] Md. M. Islam, H. Iqbal, Md. R. Haque, and Md. K. Hasan, “Prediction of breast cancer using support vector machine and K-Nearest neighbors,” in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dec. 2017, pp. 226–229. doi: 10.1109/R10-HTC.2017.8288944.
[298] Md. R. Haque, Md. M. Islam, H. Iqbal, Md. S. Reza, and Md. K. Hasan, “Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder,” in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, Feb. 2018, pp. 1–5. doi: 10.1109/IC4ME2.2018.8465658.
[299] H. Saoud, A. Ghadi, M. Ghailani, and B. A. Abdelhakim, “Application of Data Mining Classification Algorithms for Breast Cancer Diagnosis,” in *Proceedings of the 3rd International Conference on Smart City Applications*, in SCA ’18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 1–7. doi: 10.1145/3286606.3286861.
[300] K. Juneja and C. Rana, “An improved weighted decision tree approach for breast cancer prediction,” *Int. J. Inf. Technol.*, vol. 12, no. 3, pp. 797–804, Sep. 2020, doi: 10.1007/s41870-018-0184-2.
[301] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis,” *Designs*, vol. 2, no. 2, Art. no. 2, Jun. 2018, doi: 10.3390/designs2020013.
[302] B. B. A and P. Thirumalaikolundusubramanian, “Comparison of Bayes Classifiers for Breast Cancer Classification,” *Asian Pac. J. Cancer Prev. APJCP*, vol. 19, no. 10, pp. 2917–2920, 2018, doi: 10.22034/APJCP.2018.19.10.2917.
[303] A. T. Azar and S. M. El-Metwally, “Decision tree classifiers for automated medical diagnosis,” *Neural Comput. Appl.*, vol. 23, no. 7, pp. 2387–2403, Dec. 2013, doi: 10.1007/s00521-012-1196-7.
[304] M. R. Senapati, A. K. Mohanty, S. Dash, and P. K. Dash, “Local linear wavelet neural network for breast cancer recognition,” *Neural Comput. Appl.*, vol. 22, no. 1, pp. 125–131, Jan. 2013, doi: 10.1007/s00521-011-0670-y.

[305] M. R. Senapati, G. Panda, and P. K. Dash, “Hybrid approach using KPSO and RLS for RBFNN design for breast cancer detection,” *Neural Comput. Appl.*, vol. 24, no. 3, pp. 745–753, Mar. 2014, doi: 10.1007/s00521-012-1286-6.

[306] Md. K. Hasan, Md. M. Islam, and M. M. A. Hashem, “Mathematical model development to detect breast cancer using multigene genetic programming,” in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, May 2016, pp. 574–579. doi: 10.1109/ICIEV.2016.7760068.

[307] A. T. Azar and S. A. El-Said, “Performance analysis of support vector machines classifiers in breast cancer mammography recognition,” *Neural Comput. Appl.*, vol. 24, no. 5, pp. 1163–1177, Apr. 2014, doi: 10.1007/s00521-012-1324-4.

[308] P. Ferreira, I. Dutra, R. Salvini, and E. Burnside, “Interpretable models to predict Breast Cancer,” in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2016, pp. 1507–1511. doi: 10.1109/BIBM.2016.7822745.

[309] S. Jhajharia, S. Verma, and R. Kumar, “A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data,” in *2016 International Conference on Inventive Computation Technologies (ICICT)*, Aug. 2016, pp. 1–7. doi: 10.1109/INVENTIVE.2016.7830107.

[310] V. Chaurasia, S. Pal, and B. Tiwari, “Prediction of benign and malignant breast cancer using data mining techniques,” *J. Algorithms Comput. Technol.*, vol. 12, no. 2, pp. 119–126, Jun. 2018, doi: 10.1177/1748301818756225.

[311] G. Zorluoglu and M. Agaoglu, “Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods,” *Int. J. Oncol. Cancer Ther.*, vol. 02, Dec. 2017, Accessed: Nov. 13, 2022. [Online]. Available: https://www.iaras.org/iaras/home/caijoct/diagnosis-of-breast-cancer-using-ensemble-of-data-mining-classification-methods

[312] V. Chaurasia and S. Pal, “Prediction of Benign and Malignant Breast Cancer Using Data Mining Techniques.” Rochester, NY, Feb. 20, 2018. doi: 10.2139/ssrn.3139141.

[313] V. Kumar, B. K. Mishra, M. Mazzara, D. N. H. Thanh, and A. Verma, “Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications,” in *Advances in Data Science and Management*, S. Borah, V. Emilia Balas, and Z. Polkowski, Eds., in Lecture Notes on Data Engineering and Communications Technologies. Singapore: Springer, 2020, pp. 435–442. doi: 10.1007/978-981-15-0978-0_43.

[314] U. Ojha and S. Goel, “A study on prediction of breast cancer recurrence using data mining techniques,” in *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, Jan. 2017, pp. 527–530. doi: 10.1109/CONFLUENCE.2017.7943207.

[315] J. M. Rodríguez-Jiménez, P. Cordero, M. Enciso, and A. Mora, “Data mining algorithms to compute mixed concepts with negative attributes: an application to breast cancer data analysis,” *Math. Methods Appl. Sci.*, vol. 39, no. 16, pp. 4829–4845, 2016, doi: 10.1002/mma.3814.

[316] N. Al-Azzam and I. Shatnawi, “Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer,” *Ann. Med. Surg.*, vol. 62, pp. 53–64, Feb. 2021, doi: 10.1016/j.amsu.2020.12.043.

[317] K. Rajendran, M. Jayabalan, and V. Thiruchelvam, “Predicting Breast Cancer via Supervised Machine Learning Methods on Class Imbalanced Data,” *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 11, no. 8, Art. no. 8, 31 2020, doi: 10.14569/IJACSA.2020.0110808.

[318] M. Alshammari and M. Mezher, “A Comparative Analysis of Data Mining Techniques on Breast Cancer Diagnosis Data using WEKA Toolbox,” *Int. J. Adv.*

Comput. Sci. Appl. IJACSA, vol. 11, no. 8, Art. no. 8, 31 2020, doi: 10.14569/IJACSA.2020.0110829.

[319] Z. Rustam, Y. Amalia, S. Hartini, and G. S. Saragih, "Linear discriminant analysis and support vector machines for classifying breast cancer," *IAES Int. J. Artif. Intell. IJ-AI*, vol. 10, no. 1, Art. no. 1, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp253-256.

[320] W. N. L. W. H. Ibeni, M. Z. M. Salikon, A. Mustapha, S. A. Daud, and M. N. M. Salleh, "Comparative analysis on bayesian classification for breast cancer problem," *Bull. Electr. Eng. Inform.*, vol. 8, no. 4, Art. no. 4, Dec. 2019, doi: 10.11591/eei.v8i4.1628.

[321] F. S. Khan, M. I. Abbasi, M. Khurram, M. N. H. Mohd, and M. D. Khan, "Breast cancer histological images nuclei segmentation and optimized classification with deep learning," *Int. J. Electr. Comput. Eng. IJECE*, vol. 12, no. 4, Art. no. 4, Aug. 2022, doi: 10.11591/ijece.v12i4.pp4099-4110.

[322] T. S. Lim, K. G. Tay, A. Huong, and X. Y. Lim, "Breast cancer diagnosis system using hybrid support vector machine-artificial neural network," *Int. J. Electr. Comput. Eng. IJECE*, vol. 11, no. 4, Art. no. 4, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3059-3069.

[323] S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," in *2021 International Conference on Artificial Intelligence (ICAI)*, Apr. 2021, pp. 97–101. doi: 10.1109/ICAI52203.2021.9445249.

[324] A. Atrey, N. Narayan, S. Vijh, and S. Kumar, "Analysis of Breast Cancer using Machine Learning Methods," in *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Jan. 2022, pp. 258–261. doi: 10.1109/Confluence52989.2022.9734184.

[325] S. Jain and P. Kumar, "Prediction of Breast Cancer Using Machine Learning," *Recent Adv. Comput. Sci. Commun.*, vol. 13, no. 5, pp. 901–908, doi: 10.2174/2213275912666190617160834.

[326] P. P. Sengar, M. J. Gaikwad, and A. S. Nagdive, "Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Aug. 2020, pp. 796–801. doi: 10.1109/ICSSIT48917.2020.9214267.

[327] O. Tarawneh, M. Otair, M. Husni, H. Y. Abuaddous, M. Tarawneh, and M. A. Almomani, "Breast Cancer Classification using Decision Tree Algorithms," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 13, no. 4, Art. no. 4, 30 2022, doi: 10.14569/IJACSA.2022.0130478.

[328] E. Sugiharti, R. Arifudin, D. T. Wiyanti, and A. B. Susilo, "Integration of convolutional neural network and extreme gradient boosting for breast cancer detection," *Bull. Electr. Eng. Inform.*, vol. 11, no. 2, Art. no. 2, Apr. 2022, doi: 10.11591/eei.v11i2.3562.

[329] C. Kaul and N. Sharma, "High Accuracy Predictive Model on Breast Cancer Using Ensemble Approach of Supervised Machine Learning Algorithms," in *2021 International Conference on Computational Performance Evaluation (ComPE)*, Dec. 2021, pp. 071–076. doi: 10.1109/ComPE53109.2021.9752254.

[330] Md. M. Islam, Md. R. Haque, H. Iqbal, Md. M. Hasan, M. Hasan, and M. N. Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 5, p. 290, Sep. 2020, doi: 10.1007/s42979-020-00305-w.