Vol 3, Issue 12, December 2023 www.ijesti.com E-ISSN: 2582-9734
International Journal of Engineering, Science, Technology and Innovation (IJESTI)

A Study Toward the Methodology Used in Cancer Prediction Using Data Mining Techniques and Machine Learning Approaches for Early Diagnosis and Prognosis

[1] **Vimmi Kochher**

Ph.D. Scholar, School of Computer Science and Engineering
Starex University, Gurugram

v.kochher1990@yahoo.com

[2] **Dr. Shivani Sharma**

Assistant Professor
School of Computer Science and Engineering, Amity University, Gurugram

shivanijoon333@gmail.com

ABSTRACT

Cancer remains one of the leading causes of mortality worldwide, necessitating advanced techniques for early detection and accurate diagnosis. Traditional diagnostic methods are often time-consuming and expensive. This study explores the application of data mining techniques in cancer prediction, utilizing machine learning algorithms such as decision trees, support vector machines (SVM), artificial neural networks (ANN), and clustering approaches. Through analysing large medical datasets, these techniques help in identifying high-risk individuals and improving diagnostic precision. The findings suggest that data mining significantly enhances early detection, treatment planning, and survival prediction. Despite challenges like data privacy and class imbalances, advancements in artificial intelligence continue to refine these predictive models, making them more reliable and accurate. This study highlights the effectiveness, challenges, and future scope of data mining in cancer prediction, ultimately contributing to improved healthcare outcomes.

Keywords: Cancer Prediction, Data Mining, Machine Learning, Early Diagnosis, Prognosis

I. Introduction

Cancer is one of the leading causes of mortality worldwide, with millions of new cases reported annually. Early detection and accurate diagnosis are essential for improving survival rates and treatment outcomes. Traditional diagnostic methods, including biopsies and imaging techniques, are often time-consuming and expensive. As a result, researchers have turned to data mining techniques to analyse large datasets and identify patterns that can aid in early cancer prediction ***(Sarvestani et al., 2010).*** Data mining involves extracting meaningful patterns from complex datasets and applying various machine learning algorithms to classify and predict diseases. In cancer prediction, data mining techniques utilize patient records, genetic information, tumour characteristics, and medical imaging data to identify high-risk individuals ***(Fan, Zhu, & Yin, 2010).*** These techniques help healthcare professionals make more precise and timely decisions, reducing the risk of misdiagnosis and unnecessary treatments ***(Chauhan, Kaur, & Alam, 2010).*** Several machines learning methods, including decision trees, support vector machines (SVM), artificial neural networks (ANN), and deep learning approaches, have been successfully applied to cancer diagnosis and prognosis ***(Gupta, Kumar, & Sharma, 2011).*** These techniques analyse factors such as age, genetic

https://doi.org/10.31426/ijesti.2023.3.12.3951

predisposition, tumour size, and cell morphology to determine the likelihood of cancer presence. Clustering algorithms further aid in categorizing different types of cancers based on similarities in patient data ***(Agrawal & Choudhary, 2011).*** Despite its potential, cancer prediction using data mining faces challenges such as data privacy concerns, class imbalances, and the need for high-quality datasets ***(Agrawal et al., 2011).*** However, advancements in artificial intelligence, big data analytics, and cloud computing are addressing these limitations, making predictive models more accurate and reliable ***(Shouman, Turner, & Stocker, 2012).*** This study explores various data mining techniques for cancer prediction, highlighting their effectiveness, challenges, and future scope in medical research. Through leveraging machine learning, healthcare systems can improve early detection, enhance patient care, and ultimately reduce cancer-related mortality ***(Prasanna, Govinda, & Kumaran, 2012; Jacob & Ramani, 2012).***

II. Related Reviews

Sarvestani et al. (2010) aimed to identify efficient neural network architectures for analysing breast cancer data collected from clinical sources. The objective was to determine the proportion of disease progression and aid in selecting suitable treatment plans using data mining techniques. The methodology involved evaluating various neural network models, including the self-organizing map (SOM), radial basis function network (RBF), general regression neural network (GRNN), and probabilistic neural network (PNN), on datasets like the Wisconsin Breast Cancer Data (WBCD) and the Shiraz Namazi Hospital Breast Cancer Data (NHBCD). Principal component analysis was applied to reduce the dimensionality of the data and identify relevant network structures, addressing challenges posed by high-dimensional datasets. The findings revealed that the proposed networks performed effectively in analysing breast cancer data, with each model demonstrating varying degrees of accuracy and efficiency. This comparison highlighted the potential of neural networks to handle complex medical data and support clinical decision-making processes. The results were considered valuable for improving diagnostic precision and treatment selection for breast cancer patients.

Fan, Q., Zhu, C. J., & Yin, L. (2010) predict the recurrence of breast cancer using data mining techniques, utilizing the SEER Public-Use Data from 2005. The methodology involved identifying relevant information on breast cancer recurrence within the SEER dataset and employing a novel pre-classification approach for data preparation. After preparing the dataset, various data mining techniques were explored to assess their performance. Among the techniques tested, the C5 algorithm demonstrated the highest accuracy in predicting breast cancer recurrence. The findings highlighted the effectiveness of the C5 algorithm, showcasing its potential in accurately predicting cancer recurrence based on historical data. This study's relevance lies in its contribution to improving predictive models in cancer research, offering valuable insights into the application of data mining for early detection and recurrence prediction in breast cancer, ultimately aiding in personalized treatment strategies and timely interventions.

Chauhan, Kaur, and Alam (2010) aimed to explore the application of data mining techniques and tools for analysing medical and geographical data. The methodology involved using clustering techniques, such as hierarchical and classical clustering, to analyse both discrete and continuous spatial medical datasets. The researchers applied these methods to geographical data to discover valuable patterns, focusing on generating efficient clusters from spatial data. The findings revealed that clustering methods could effectively uncover meaningful insights, especially when applied to continuous datasets, which required clusters of arbitrary shapes for better results. Additionally, the study highlighted the importance of deep analysis, as some facts developed over time were not easily extracted from raw data through quick or superficial means. This research contributes significantly to the field of spatial data mining by demonstrating the effectiveness of clustering techniques in discovering hidden patterns in large datasets, offering valuable insights for medical and geographical data analysis. The study is highly relevant for advancing data mining applications in complex data environments where traditional methods may fall short.

Gupta, Kumar, and Sharma (2011) aimed to explore how machine learning and data mining techniques have transformed the diagnosis and prognosis of breast cancer. The research focused on distinguishing between benign and malignant breast lumps (diagnosis) and predicting the likelihood of recurrence in patients who had undergone tumour removal (prognosis). The methodology involved reviewing a range of technical publications and research focused on applying data mining methods to these medical challenges. The findings highlighted how these techniques have significantly improved both the accuracy and efficiency of breast cancer diagnosis and prognosis. The study emphasized the relevance of data mining in the medical field, particularly in categorizing cancer cases and predicting outcomes, which has the potential to enhance early detection and treatment strategies. This research contributed to the growing body of knowledge on the application of data mining in healthcare, providing valuable insights into how these methods can be leveraged for more effective breast cancer management.

Agrawal and Choudhary (2011) conducted a study using association rule mining techniques to analyze lung cancer data from the SEER program, aiming to identify patient subgroups with significantly higher or lower survival rates. Their objective was to locate hotspots within the data by evaluating thirteen patient characteristics linked to survival outcomes. The methodology involved generating numerous association rules, which were manually filtered to remove redundant ones based on domain expertise. These rules were then used for segmentation of the dataset, focusing on identifying subgroups where the survival rates deviated substantially from the overall average. The findings revealed that the criteria developed aligned with existing scientific knowledge and provided valuable insights into lung cancer patient survival, confirming certain characteristics that could influence prognosis. The study's relevance lies in its application of data mining techniques to improve understanding of cancer survival, offering a foundation for further research in predictive modelling and patient subgroup identification.

Agrawal et al. (2011) conducted an analysis of lung cancer data provided by the SEER program to develop accurate survival prediction models using data mining techniques. The study's objective was to identify predictive attributes and construct a reliable model for forecasting lung cancer outcomes. The methodology involved preprocessing the data by removing, modifying, or splitting various attributes to enhance the quality of the dataset. After preprocessing, they applied several data mining classification methods along with optimizations and validations. The study found that ensemble voting of five decision tree-based classifiers and meta-classifiers achieved the highest prediction performance in terms of accuracy and area under the ROC curve. Additionally, the researchers developed an online lung cancer outcome calculator to estimate mortality risk at different time intervals (6 months, 9 months, 1 year, 2 years, and 5 years) post-diagnosis. A subset of thirteen non-redundant attributes was selected using attribute selection techniques, ensuring that the predictive power of the original dataset was preserved. This study was relevant for advancing cancer prognosis by integrating data mining into clinical applications, helping in risk prediction and aiding medical professionals in decision-making processes.

Shouman, Turner, and Stocker (2012) explore the use of data mining and statistical techniques in the diagnosis and treatment of heart disease. The methodology involved analysing large medical datasets and evaluating the effectiveness of data mining approaches, specifically focusing on the use of hybrid techniques to improve diagnostic accuracy. The findings revealed that while single data mining approaches yielded acceptable accuracy levels in detecting heart disease, the hybridization of multiple techniques showed enhanced performance in diagnosis. However, the study also highlighted a gap in research concerning the use of data mining methods for determining appropriate treatments for heart disease. The authors proposed a model aimed at bridging this gap, hypothesizing that the application of data mining techniques in treatment decision-making could achieve similar performance to that of diagnosis. The study was relevant as it addressed the growing need for efficient data analysis tools in the medical field, particularly in the context of heart disease, and provided insights into improving both diagnostic and therapeutic approaches through advanced data mining methods.

Prasanna, S., Govinda, K., and Kumaran, U. S. (2012) aimed to address the growing challenge of identifying oral cancer, which is particularly prevalent in developing South Asian nations like India. The authors focused on implementing two prominent data mining techniques Naïve Bayesian and Support Vector Machine (SVM) to predict cancer based on patient data. The methodology involved creating the Intelligent Prognosis Prediction System for Cancer Disease (IPPSCD), a database used to store cancer patient information. The results of both methods were compared to identify the more effective approach for predicting the disease. Their findings suggested that data mining techniques, particularly SVM, provided more accurate predictions compared to traditional linear and logistic regression models commonly used in medical oncology. This study highlighted the relevance of advanced computational techniques in improving early cancer detection and prognostic analysis, contributing valuable insights to the medical field, especially in regions with high oral cancer rates.

Jacob and Ramani (2012) aimed to develop an effective classifier for analysing data related to the status of breast cancer patients, focusing on utilizing data mining techniques for clinical studies. The researchers designed a framework that integrates pattern learning and rule extraction to enhance decision-making in new situations. They employed machine learning methods, including feature relevance analysis and classification algorithms, to train the system using data from the Wisconsin Prognostic Breast Cancer (WPBC) repository. The dataset comprised 198 instances with 33 predictor variables, and the study primarily evaluated the impact of feature reduction and classification methods. The Random Tree and C4.5 algorithms were used, with specific parameters such as the number of features and confidence levels being optimized. The findings showed that, with appropriate parameter selection, the classifiers achieved a perfect accuracy rate of 100% in both the training and testing phases. The study demonstrated the potential of machine learning methods in achieving high classification accuracy for breast cancer patient status, highlighting the relevance of data mining in medical diagnostics.

Kolçe and Frasheri (2012) aimed to provide an overview of the data mining methods applied in diagnosing and predicting outcomes for various illnesses, with a focus on identifying the most effective algorithms for medical datasets. The researchers explored several algorithms, including Decision Trees, Support Vector Machines, Artificial Neural Networks (specifically the Multilayer Perceptron model), Naïve Bayes, and Fuzzy Rules. The methodology involved conducting various analyses to assess the performance of these methods. The findings indicated that no single data mining technique could be deemed universally superior for illness diagnosis or prognosis. While some algorithms showed better performance in specific cases, combining the advantageous features of multiple algorithms led to more impressive results in certain scenarios. The study highlighted the complexity of selecting a single method for medical predictions, suggesting that hybrid approaches may offer superior accuracy. This research is relevant as it provides valuable insights into the application of data mining in healthcare, showing that a tailored approach incorporating multiple techniques could enhance predictive accuracy in medical diagnostics and prognosis.

Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013) aimed to explore the use of classification-based data mining techniques for the early detection and accurate diagnosis of lung cancer. The methodology involved applying various data mining approaches, such as Rule-based, Decision Tree, Naïve Bayes, and Artificial Neural Networks, to healthcare data. Additionally, the research utilized advanced classifiers like the One Dependency Augmented Naïve Bayes (ODANB) and the naïve creedal classifier 2 (NCC2) to handle incomplete or imprecise data. The study found that these methods could successfully detect lung cancer at an early stage, which is crucial for preventing unnecessary complications and improving patient survival. It highlighted the importance of identifying hidden patterns in large healthcare datasets, which are often overlooked despite the vast amounts of data collected. The study’s relevance lies in its contribution to the healthcare sector, where early detection models can significantly enhance decision-making and help in reducing mortality rates from lung cancer. The research demonstrated that using general symptoms, such as age, gender, and physical symptoms like shortness of breath, could offer insights into a patient’s likelihood of developing lung cancer, thus helping physicians provide timely interventions.

Vijiyarani and Sudha (2013) explore the application of data mining techniques in predicting diseases such as heart disease, diabetes, and breast cancer, aiming to reduce the number of diagnostic tests required for disease detection. The methodology involved reviewing various research publications on data mining methods, including association rules, classification, clustering, prediction, and sequential patterns, which are commonly used in health care for disease prediction. The findings highlighted the effectiveness of data mining in minimizing the number of tests while improving prediction accuracy, thus enhancing both time efficiency and performance in diagnosing diseases. The study also acknowledged the pros and cons of utilizing data mining methods, noting their significant role in the healthcare industry while also addressing potential drawbacks. The relevance of this study lies in its contribution to the healthcare sector by demonstrating how data mining can streamline diagnostic processes and improve disease prediction, ultimately aiding in better patient care and resource management.

Chandrasekar, R. M., Palaniammal, V., & Phil, M. (2013) aimed to analyse breast cancer data using various data mining methods to explore the effectiveness of different classification techniques. The dataset, sourced from the World Breast Cancer Organization (WBC) and the Women's Breast Cancer Centre (WDBC) at the University of California, Irvine (UCI), contained 286 rows and 10 columns. The study employed several classification algorithms, including Decision Tree, Rules NNge, Random Forest, Random Tree, and IBK, within the WEKA environment. The researchers focused on comparing the performance of these algorithms in predicting breast cancer outcomes. The results revealed that different classification methods exhibited varying levels of accuracy, with some models outperforming others in terms of precision and recall. The findings emphasized the potential of data mining techniques in the early detection and classification of breast cancer, offering valuable insights for future research in both scientific and commercial applications. This study highlighted the growing relevance of data mining in healthcare, showcasing its ability to contribute to critical decision-making in disease diagnosis.

Majali et al. (2014) aimed to develop a system for early cancer diagnosis and prediction, focusing on breast cancer, which is a leading cause of mortality in women worldwide. The methodology involved using the Frequent Pattern (FP) growth algorithm to analyse cancer data, along with the Decision Tree method to predict the likelihood of cancer development based on age. The FP algorithm was specifically employed to determine whether tumours were malignant or benign. The study found that early detection through these data mining techniques could significantly improve prognosis, with 97% of women surviving for five or more years if diagnosed early. The relevance of this research lies in its potential to enhance diagnostic capabilities, particularly in developing countries like India, where breast cancer is increasingly prevalent. Through data mining, the study proposed a more efficient approach to diagnosing cancer at an earlier stage, which could lead to better treatment outcomes and reduced mortality.

Wang and Yoon (2015) aimed to identify a reliable strategy for predicting breast cancer using data mining techniques. The study compared various models, including support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, and AdaBoost tree, to find an accurate model for forecasting breast cancer based on clinical data. The research emphasized the significant impact of feature space on the learning process, proposing a hybrid approach combining principal component analysis (PCA) with the data mining models to reduce feature space. The study used two widely recognized test datasets, the Wisconsin Breast Cancer Database (1991) and the Wisconsin Diagnostic Breast Cancer (1995), and employed 10-fold cross-validation to estimate the test error for each model. The findings provided a comprehensive evaluation of the models, highlighting a trade-off between their performance. This research is relevant as it offers insights into the practical application of data mining in breast cancer prediction, with potential benefits for both patients and clinicians by aiding in early detection and prevention.

Bahrami and Shirvani (2015) aimed to assess various classification methods for diagnosing heart disease using data mining techniques. They explored the use of classifiers such as J48 Decision Tree, K Nearest Neighbours (KNN), Naive Bayes (NB), and SMO to analyse heart disease datasets. The study involved evaluating the classifiers based on performance metrics like accuracy, precision, sensitivity, specificity, F-measure, and the area under the ROC curve. The results indicated that the J48 Decision Tree classifier outperformed others in terms of effectiveness for heart disease diagnosis. The study emphasized the importance of data mining methods in the healthcare sector, particularly in heart disease detection, and suggested that clinicians could benefit from these tools to identify hidden patterns within large medical datasets. This research is relevant as it demonstrates the potential of data mining to improve diagnostic accuracy, which could lead to more timely and effective healthcare interventions.

Zand (2015) aimed to provide a comparative review of data mining techniques in the diagnosis and prognosis of breast cancer, specifically focusing on predicting the survival rate of patients using the SEER Public Use Data. The study employed various data mining approaches, including classification, regression, and clustering techniques, to analyse the dataset. Through examining factors such as tumour size, patient age, and other clinical variables, the research sought to identify patterns that could predict survival outcomes and improve diagnostic accuracy. The findings revealed that data mining methods, particularly decision trees and support vector machines, could accurately predict patient survival, offering valuable insights into the genetic behaviour of tumours and enabling better-targeted treatments. The study emphasized the importance of utilizing such techniques to enhance early detection and personalized medicine. This work was highly relevant as it contributed to the growing body of knowledge on leveraging data mining for improving breast cancer prognosis, a critical issue given the global prevalence and mortality rate of the disease.

Shukla, Gupta, and Prasad (2016) examine the significance and application of data mining techniques in the healthcare sector, specifically focusing on predicting cancer illness using clinical datasets. The methodology involved reviewing various data mining methods such as classification, clustering, Decision Tree, and Naive Bayes, which have shown promise in revealing hidden patterns within healthcare data. The study highlighted how these techniques have been successfully employed to analyse large datasets, enabling the prediction of cancer with varying degrees of accuracy. The findings emphasized that advancements in data mining have made it a critical tool in healthcare, offering valuable insights into patient data and aiding in decision-making. The study's relevance lies in its contribution to understanding how data mining can be leveraged to improve healthcare outcomes, particularly in the early detection and prediction of cancer, thereby enhancing the potential for timely interventions.

Assari, Azimi, and Taghva's (2017) study was to enhance early detection and prediction of cardiovascular disease by using data mining techniques. They aimed to identify key diagnostic indicators and construct a model for better heart disease diagnosis. The methodology involved consulting specialists to determine primary diagnostic indicators for heart disease, followed by the application of data mining methods to a heart-related dataset. The researchers used Visual Studio to develop the algorithm for model construction. The findings revealed that various data mining strategies applied to different datasets yielded different outcomes, underscoring the importance of selecting appropriate techniques. Ultimately, the study identified the primary indices for diagnosing heart disease and developed a model based on the extracted rules. This research is highly relevant as it provides medical professionals with tools to detect heart disease at an earlier stage, assess associated risk factors, and potentially reduce treatment costs.

Almarabeh and Amer (2017) conducted a study to explore the application of data mining techniques in healthcare, aiming to identify the most accurate and reliable prediction methods for various medical conditions. The objective of their research was to provide an overview of how data mining is being utilized across different fields, such as heart disease, diabetes, breast and lung cancer, and skin diseases. The methodology involved analysing large datasets to uncover patterns that could be used for predicting the likelihood of future occurrences of these diseases. The study found that data mining techniques have great potential in improving prediction accuracy, allowing for more effective early detection and treatment. The study is highly relevant as it highlights the increasing role of data mining in healthcare, offering valuable insights into how these methods can be applied to enhance medical decision-making and patient care.

Tafish and El-Halees (2018) develop a data mining model to predict the severity of breast cancer and assist in the diagnosis process. The methodology involved collecting breast cancer data from hospitals in the Gaza Strip and applying various classification techniques, including Support Vector Machines, Artificial Neural Networks, and k-Nearest Neighbours, alongside association rule mining to identify key characteristics linked to high-severity cases. The model aimed to improve the accuracy of breast cancer diagnosis while reducing the associated time and costs. The findings revealed that the model achieved an accuracy rate of 77%, which was considered acceptable for predicting the severity of breast cancer. It also highlighted several prominent features that were most strongly associated with severe breast cancer cases. This study is relevant to the field of medical data mining as it demonstrates the potential of machine learning algorithms to enhance diagnostic accuracy and efficiency, providing valuable insights for healthcare professionals.

Kaur and Bawa (2018) aimed to highlight the significance of using open-source data mining suites in extracting valuable insights from large datasets, particularly in the medical field, where the prevalence of diseases is a major concern. Their methodology focused on examining various data mining techniques to improve clinical decision-making and extend patient survival by accurately diagnosing illnesses. The researchers emphasized the necessity of developing new methods due to the vast complexity of data collection in fields like healthcare, administration, and commerce. They noted that while traditional diagnoses relied heavily on a doctor's expertise, errors still occurred, making data mining a crucial tool in reducing misdiagnoses. The findings revealed that selecting the appropriate data mining approach is critical for ensuring accuracy in predictions, especially in diagnosing incurable diseases, which are both costly and life-threatening. This research underlined the importance of data mining in enhancing the efficiency of healthcare systems, potentially saving lives and reducing healthcare costs. The study's relevance is clear in the context of modern medical practices, where data-driven decision-making is essential for improving patient outcomes and combating life-threatening diseases.

Mia et al. (2018) explore and identify various data mining techniques and approaches used in healthcare research, specifically focusing on their application in medical data analysis. The methodology involved reviewing academic literature to assess the different data mining methods employed for disease analysis, with an emphasis on determining their accuracy in diagnosing and predicting various illnesses. Several techniques were analyzed across multiple disease categories to evaluate their effectiveness. The study's findings revealed that no single data mining approach could comprehensively address all types of disease analysis, as each method exhibited varying degrees of accuracy depending on the illness being examined. The research underscored the complexity and challenges of extracting meaningful insights from medical data and highlighted the importance of selecting appropriate data mining tools for specific healthcare needs. This study is highly relevant as it provides valuable insights for healthcare professionals, helping them make informed decisions when choosing data mining technologies for analysing medical data.

Eltalhi and Kutrani (2019) aimed to investigate the role of machine learning and data mining techniques in the early detection and diagnosis of breast cancer, a leading cause of mortality among women. The study primarily focused on evaluating various classification algorithms for breast cancer prediction using the WEKA tool. Several algorithms, including Decision Tree, Naïve Bayes, and Artificial Neural Networks, were compared for their effectiveness in diagnosing breast cancer and predicting its progression. The methodology involved the application of data mining techniques to historical cancer data, enabling the identification of patterns that could predict the disease at an early stage. The findings indicated that machine learning and data mining approaches significantly improved breast cancer diagnosis by providing reliable predictions and aiding in early-stage detection, which is crucial for improving patient survival rates. The study highlighted the importance of employing advanced computational techniques, such as classification algorithms, in clinical settings for breast cancer detection. This research is relevant as it underscores the potential of data mining in healthcare, particularly in improving breast cancer diagnosis and supporting timely interventions.

Mabu, A. M., Prasad, R., and Yadav, R. (2020) aimed to explore the application of data mining techniques, particularly clustering and classification, in the analysis of gene expression datasets for cancer diagnosis and prognosis. Their methodology involved examining various supervised and unsupervised algorithms used in the clustering and classification of gene expression data, focusing on their relevance and efficiency in detecting malignancies. The study highlighted the significance of clustering as an unsupervised learning technique, which groups data based on similarities and distances, and classification, which assigns data to specific categories for accurate prediction. The authors analyzed the current data mining approaches in the context of gene expression data, discussing the suitability of different algorithms for identifying cancerous conditions. Their findings indicated that data mining techniques, particularly clustering, are powerful tools for detecting malignancies from gene expression data, offering promising potential in medical diagnoses. This study is highly relevant for the field of bioinformatics, where data mining methods can significantly enhance the precision of cancer diagnosis and prognosis by analysing large and complex datasets.

Biwalkar, Gupta, and Dharadhar (2021) aimed to explore the application of data mining tools within the healthcare sector, focusing on their adaptability, reliability, and the absence of limitations on data types. The authors reviewed previous research in the field and proposed a framework to enhance the accuracy of healthcare-related predictions. The methodology involved examining various data mining techniques and statistical tests that could be applied to healthcare datasets, such as for client relationship management, resource allocation, fraud detection, and treatment improvement. The findings indicated that data mining techniques have shown significant potential in enhancing the management and diagnostic capabilities within healthcare, offering reliable and efficient solutions to complex issues. The study's relevance lies in its contribution to the growing body of research on improving healthcare outcomes through data mining, which can lead to more accurate diagnoses, better resource management, and overall enhancements in healthcare delivery.

Arivazhagan et al. (2022) aimed to improve the early detection of skin cancer by proposing an advanced image analysis technique for skin lesion classification. Their methodology involved the use of cutting-edge image processing technologies to analyse skin lesions, focusing on features such as asymmetry, borders, pigment, and diameter. The researchers employed lesion image analysis tools to evaluate texture, size, and shape, which helped in segmenting images and identifying key features indicative of skin cancer. Their findings revealed that the use of these advanced techniques could significantly enhance the accuracy of skin cancer detection, thus increasing the likelihood of early diagnosis and better patient outcomes. Furthermore, the study highlighted the importance of reducing UV exposure to lower skin cancer risk, emphasizing the direct relationship between genetic mutations caused by UV radiation and the development of skin cancer. The study's relevance lies in its potential to aid in the early discovery of skin cancer, which could lead to higher survival rates due to timely intervention and the advancements in cancer treatment and diagnostic technologies.

Patela (2023) aimed to explore the use of data mining techniques in the early detection of breast cancer, emphasizing the importance of early diagnosis to reduce mortality rates. The study involved analysing large datasets, such as the Wisconsin Breast Cancer Dataset (WBCD), to uncover valuable insights that could aid in breast cancer diagnosis. The methodology included the application of various data mining classifiers, with a focus on classification accuracy to evaluate their performance. The research compared different techniques, providing a comprehensive review of their effectiveness in predicting breast cancer outcomes. The findings revealed that data mining strategies could significantly enhance diagnostic accuracy, with some classifiers outperforming others in terms of predictive power. The study highlighted the potential of data mining in healthcare, particularly for early cancer detection, and reinforced its relevance in the ongoing efforts to reduce cancer-related deaths. Through summarizing various data mining approaches, the research contributed valuable knowledge to the scientific community and emphasized the promising role of these technologies in improving healthcare outcomes. This study was highly relevant, given the global prevalence of breast cancer and the need for more efficient diagnostic tools.

III. Various Reviews and Findings

Author(s) and Year	Study Objective	Methodology	Findings	Relevance
Sarvestani et al. (2010)	Identify efficient neural network architectures for analysing breast cancer data.	Evaluated various neural networks like SOM, RBF, GRNN, PNN on WBCD & NHBCD datasets.	Neural networks effectively analyse breast cancer data.	Supports clinical decision-making in breast cancer treatment.
Fan, Zhu, & Yin (2010)	Predict breast cancer recurrence using data mining techniques.	Used SEER dataset with C5 algorithm for highest accuracy in predicting recurrence.	C5 algorithm demonstrated highest accuracy for recurrence prediction.	Enhances predictive modelling for breast cancer recurrence.
Chauhan, Kaur, & Alam (2010)	Apply clustering techniques for analysing spatial medical datasets.	Applied hierarchical and classical clustering for spatial data analysis.	Clustering methods effectively uncover insights in spatial data.	Contributes to spatial data mining in medical analysis.
Gupta, Kumar, & Sharma (2011)	Explore machine learning techniques for breast cancer diagnosis and prognosis.	Reviewed research on data mining techniques in breast cancer diagnosis and prognosis.	Data mining significantly improves breast cancer diagnosis and prognosis.	Improves early detection and treatment strategies for cancer.
Agrawal & Choudhary (2011)	Use association rule mining to identify lung cancer patient subgroups.	Generated and filtered association rules for lung cancer patient segmentation.	Association rules align with scientific knowledge on lung cancer survival.	Provides a foundation for further predictive modelling research.
Agrawal et al. (2011)	Develop lung cancer survival prediction models using data mining.	Used ensemble voting of five classifiers for survival prediction.	Ensemble classifiers achieved highest prediction accuracy.	Advances lung cancer prognosis using data mining.
Shouman, Turner, & Stocker (2012)	Explore data mining for heart disease diagnosis and treatment.	Analysed medical datasets with hybrid data mining techniques.	Hybrid techniques enhanced diagnostic accuracy for heart disease.	Addresses gaps in heart disease treatment prediction.
Prasanna, Govinda, & Kumaran (2012)	Evaluate oral cancer detection using NaÃ¯ve Bayesian and SVM.	Compared classification and regression methods using IPPSCD database.	SVM outperformed other models in predicting oral cancer.	Improves early detection and prognostic analysis of oral cancer.
Jacob & Ramani (2012)	Develop a classifier for breast cancer patient status analysis.	Integrated pattern learning and rule extraction for classification.	Random Tree and C4.5 achieved 100% accuracy in classification.	Demonstrates high classification accuracy for breast cancer status.
KolÃ§e & Frasheri (2012)	Review data mining techniques used in healthcare databases.	Compared Decision Trees, SVM, ANN, and Fuzzy Rules for medical predictions.	Hybrid approaches provide better predictive accuracy in healthcare.	Highlights need for hybrid data mining approaches in medicine.
Krishnaiah, Narsimha, & Chandra (2013)	Apply classification-based data mining for lung cancer detection.	Applied Rule-based, Decision Tree, and NaÃ¯ve Bayes for lung cancer detection.	Advanced classifiers detected lung cancer at early stages.	Supports early detection of lung cancer.
Vijiyarani & Sudha (2013)	Explore data mining for predicting heart disease, diabetes, and breast cancer.	Reviewed research on data mining for predicting multiple diseases.	Data mining improved diagnostic efficiency for multiple diseases.	Enhances diagnosis and treatment efficiency.
Chandrasekar, Palaniammal, & Phil (2013)	Analyse breast cancer data with various classification techniques.	Compared classification algorithms like Decision Tree and Random Forest.	Different classification algorithms showed varying accuracy.	Demonstrates effectiveness of classification algorithms.

Majali et al. (2014)	Use FP-growth and Decision Tree for early breast cancer detection.	Applied FP-growth and Decision Tree for tumour classification.	FP-growth and Decision Tree improved early breast cancer detection.	Improves diagnostic capabilities in developing countries.
Wang & Yoon (2015)	Compare machine learning models for breast cancer prediction.	Evaluated SVM, ANN, Naave Bayes, and AdaBoost for breast cancer prediction.	Hybrid models showed improved breast cancer prediction accuracy.	Enhances prediction accuracy in breast cancer detection.
Bahrami & Shirvani (2015)	Evaluate classifiers for heart disease diagnosis.	Compared J48, KNN, Naave Bayes, and SMO classifiers.	J48 Decision Tree was most effective for heart disease diagnosis.	Advances heart disease diagnostic models.
Zand (2015)	Review data mining for predicting breast cancer survival.	Applied classification, regression, and clustering on SEER dataset.	Decision Trees and SVM accurately predicted patient survival.	Improves survival prediction for breast cancer patients.
Shukla, Gupta, & Prasad (2016)	Analyse data mining in predicting cancer using clinical datasets.	Reviewed data mining models for predicting cancer.	Machine learning methods enhanced cancer prediction.	Refines early cancer prediction methods.
Assari, Azimi, & Taghva (2017)	Use data mining for early cardiovascular disease detection.	Developed a cardiovascular disease prediction model.	Developed a cardiovascular disease detection model.	Helps physicians in early cardiovascular disease detection.
Almarabeh & Amer (2017)	Apply data mining for accurate disease prediction.	Used Decision Trees, SVM, ANN, and Naave Bayes for disease prediction.	Machine learning significantly improved medical predictions.	Improves disease monitoring and prediction accuracy.
Tafish & El-Halees (2018)	Predict breast cancer severity using machine learning.	Applied classification algorithms to hospital datasets.	AI models achieved 77% accuracy in cancer severity prediction.	Supports AI-driven diagnosis of severe breast cancer cases.
Kaur & Bawa (2018)	Evaluate open-source data mining tools in healthcare.	Explored open-source data mining software applications.	Open-source tools effectively analysed large medical datasets.	Improves access to medical data analysis tools.
Mia et al. (2018)	Review data mining applications in medical data analysis.	Reviewed multiple data mining techniques in medical research.	No single data mining approach was universally superior.	Encourages tailored approaches for disease prediction.
Eltalhi & Kutrani (2019)	Compare classification algorithms for breast cancer detection.	Used classification algorithms on cancer datasets.	Machine learning algorithms improved breast cancer classification.	Aids in early cancer diagnosis.
Mabu, Prasad, & Yadav (2020)	Explore clustering and classification for cancer gene expression analysis.	Analysed supervised and unsupervised learning for gene expression data.	Gene expression clustering accurately identified malignancies.	Enhances precision in genetic-based cancer detection.
Biwalkar, Gupta, & Dharadhar (2021)	Assess data mining's impact on healthcare predictions.	Reviewed machine learning for predicting diseases.	Data mining enhanced healthcare resource management.	Provides insights for disease prediction frameworks.
Arivazhagan et al. (2022)	Use image processing for early skin cancer detection.	Used advanced image processing for skin lesion classification.	Image processing improved early skin cancer diagnosis.	Improves skin cancer early diagnosis.
Patela (2023)	Apply data mining for early breast cancer detection.	Applied data mining classifiers on WBCD dataset.	Data mining significantly improved breast cancer detection rates.	Refines machine learning for medical applications.

IV. Methodology

The methodologies used in cancer prediction studies utilizing data mining techniques involve various machine learning algorithms, feature selection methods, and data preprocessing techniques. These approaches aim to extract valuable insights from medical datasets and enhance early cancer diagnosis and prognosis.

4.1 Data Collection and Preprocessing

The first step in data mining-based cancer prediction is collecting data from reliable sources such as the Wisconsin Breast Cancer Database (WBCD), Surveillance, Epidemiology, and End Results (SEER) program, and other hospital datasets. These datasets contain essential patient details, including demographics, tumour characteristics, and survival rates ***(Krishnaiah, Narsimha, & Chandra, 2013)***.
Before applying machine learning techniques, data preprocessing is crucial. This step includes handling missing values, removing duplicate entries, normalizing numerical values, and encoding categorical variables ***(Vijiyarani & Sudha, 2013)***. Feature selection is applied to reduce dimensionality and improve prediction accuracy, often utilizing Principal Component Analysis (PCA) or correlation-based feature selection methods.

4.2 Classification Algorithms for Cancer Prediction

Cancer diagnosis and prognosis largely depend on classification algorithms, which predict whether a patient has cancer or the likelihood of recurrence. Various machine learning algorithms are commonly used.

Decision Trees (C4.5, J48): Used for classification problems by creating hierarchical structures for decision-making (***Chandrasekar, Palaniammal, & Phil, 2013).***

Support Vector Machines (SVM): Efficient for high-dimensional datasets and widely used in cancer diagnosis ***(Majali et al., 2014).***

Naïve Bayes: Based on probability theory, this algorithm is useful for identifying cancer subtypes based on medical history.

k-Nearest Neighbours (KNN): Used for classifying cancer cases based on similarity with previously classified cases ***(Wang & Yoon, 2015).***

Random Forest: An ensemble method that improves classification accuracy by combining multiple decision trees.

Artificial Neural Networks (ANN): Deep learning models, including Convolutional Neural Networks (CNNs), are effective in medical imaging analysis ***(Bahrami & Shirvani, 2015).***

4.3 Clustering Techniques in Cancer Analysis

Clustering techniques are used to group similar cancer cases for better understanding:

k-Means Clustering: Groups patients into different risk categories based on tumour size and genetic markers.

Hierarchical Clustering: Used to analyse spatial and genetic data in cancer research ***(Zand, 2015).***

Self-Organizing Maps (SOMs): A type of neural network used to cluster high-dimensional cancer datasets ***(Shukla, Gupta, & Prasad, 2016).***

4.4 Association Rule Mining for Pattern Discovery

Some studies have applied association rule mining techniques such as A-priori Algorithm and FP-Growth Algorithm to find hidden relationships between cancer attributes. These methods help identify the most significant risk factors influencing cancer progression ***(Assari, Azimi, & Taghva, 2017).***

4.5 Ensemble Learning and Hybrid Approaches

To improve prediction accuracy, ensemble methods like bagging, boosting, and stacking are employed. Studies have found that ensemble learning, which combines multiple classifiers, often results in higher prediction performance ***(Almarabeh & Amer, 2017).*** Hybrid models integrating multiple data mining techniques have also been shown to enhance cancer detection efficiency.

V. Application of Data Mining in Cancer Prediction

The application of data mining techniques in cancer prediction has revolutionized the field of medical diagnostics, enabling early detection and more precise treatment strategies. These techniques analyse large volumes of medical data to identify patterns and correlations that help predict cancer presence, recurrence, and severity. Data mining has been particularly effective in breast cancer diagnosis, lung cancer prognosis, and skin cancer detection, improving patient survival rates and reducing unnecessary procedures.

Breast cancer diagnosis has significantly benefited from data mining applications. Machine learning models, such as support vector machines (SVM) and artificial neural networks (ANN), have been extensively used to classify benign and malignant tumours. Tafish and El-Halees (2018) demonstrated that applying data mining techniques, such as classification and regression, could accurately predict the severity of breast cancer cases. Similarly, Kaur and Bawa (2018) reviewed various data mining techniques used in the healthcare sector and emphasized their importance in improving early cancer detection and treatment planning. Another crucial application of data mining is in lung cancer prognosis. According to Mia, Hossain, Chhoton, and Chakraborty (2018), data mining approaches, including clustering and classification algorithms, play a vital role in predicting lung cancer outcomes by analyzing genetic and clinical datasets. Eltalhi and Kutrani (2019) further explored the potential of data mining in breast cancer prediction, highlighting that combining multiple machine learning algorithms enhances diagnostic accuracy and reliability.

Gene expression analysis is another emerging area where data mining techniques contribute to cancer research. Mabu, Prasad, and Yadav (2020) reviewed various methods used in mining gene expression data and found that clustering algorithms, such as k-means and hierarchical clustering, help in identifying significant genetic markers associated with different types of cancer. These insights are crucial for personalized medicine and targeted therapies. Skin cancer detection has also advanced with the application of data mining. Arivazhagan et al. (2022) demonstrated how computational intelligence techniques, including deep learning models and image processing, could enhance the accuracy of diagnosing skin cancer from dermoscopic images. Through utilizing convolutional neural networks (CNNs), data mining enables automated classification of skin lesions, reducing the need for invasive diagnostic procedures.

Recent advancements in data mining techniques for cancer diagnosis include the integration of hybrid models that combine multiple classifiers for improved predictive performance. Biwalkar, Gupta, and Dharadhar (2021) emphasized that hybrid approaches leveraging decision trees, random forests, and neural networks provide higher accuracy in diagnosing various types of cancer. Additionally, Patela (2023) conducted a study on breast cancer diagnosis using data mining, demonstrating the efficacy of different classifiers in improving early detection rates.

VI. Conclusion

Data mining techniques have proven to be highly effective in cancer prediction, offering improved accuracy and efficiency over traditional diagnostic methods. Through machine learning algorithms, these techniques facilitate early detection, classification, and prognosis of various cancers, including breast, lung, and oral cancer. Studies have demonstrated that methods like decision trees, SVM, ANN, and ensemble learning models significantly enhance predictive accuracy, aiding in timely treatment interventions. Additionally, clustering and association rule mining help identify patterns and risk factors within large datasets, supporting clinical decision-making. Despite their potential, data mining approaches face several challenges, including data privacy concerns, imbalanced datasets, and the need for high-quality medical data. However, advancements in artificial intelligence, deep learning, and big data analytics continue to address these limitations, making cancer prediction models more robust and interpretable. Future research should focus on integrating hybrid models and deep learning techniques to further enhance diagnostic precision.

References

1. Sarvestani, A. S., Safavi, A. A., Parandeh, N. M., & Salehi, M. (2010, October). Predicting breast cancer survivability using data mining techniques. In *2010 2nd international conference on software technology and engineering* (Vol. 2, pp. V2-227). IEEE.
2. Fan, Q., Zhu, C. J., & Yin, L. (2010, April). Predicting breast cancer recurrence using data mining techniques. In *2010 International Conference on Bioinformatics and Biomedical Technology* (pp. 310-311). IEEE.
3. Chauhan, R., Kaur, H., & Alam, M. A. (2010). Data clustering method for discovering clusters in spatial cancer databases. *International Journal of Computer Applications*, *10*(6), 9-14.
4. Gupta, S., Kumar, D., & Sharma, A. (2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering (IJCSE)*, *2*(2), 188-195.
5. Agrawal, A., & Choudhary, A. (2011, December). Identifying hotspots in lung cancer data using association rule mining. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 995-1002). IEEE.
6. Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L., & Choudhary, A. (2011, August). A lung cancer outcome calculator using ensemble data mining on SEER data. In *Proceedings of the tenth international workshop on data mining in bioinformatics* (pp. 1-9).
7. Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. In *2012 Japan-Egypt Conference on Electronics, Communications and Computers* (pp. 173-177). IEEE.
8. Prasanna, S., Govinda, K., & Kumaran, U. S. (2012). An Evaluation study of Oral Cancer Detection using Data Mining Classification Techniques. *International Journal of Advanced Research in Computer Science*, *3*(1).

9. Jacob, S. G., & Ramani, R. G. (2012, October). Efficient classifier for classification of prognostic breast cancer data through data mining techniques. In *Proceedings of the World Congress on Engineering and Computer Science* (Vol. 1, pp. 24-26).
10. Kolçe, E., & Frasheri, N. (2012). A literature review of data mining techniques used in healthcare databases. *ICT innovations*.
11. Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. *International Journal of Computer Science and Information Technologies*, *4*(1), 39-45.
12. Vijiyarani, S., & Sudha, S. (2013). Disease prediction in data mining technique–a survey. *International Journal of Computer Applications & Information Technology*, *2*(1), 17-21.
13. Chandrasekar, R. M., Palaniammal, V., & Phil, M. (2013). Performance and evaluation of data mining techniques in cancer diagnosis. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *15*(5), 39-44.
14. Majali, J., Niranjan, R., Phatak, V., & Tadakhe, O. (2014). Data mining techniques for diagnosis and prognosis of breast cancer. *International Journal of Computer Science and Information Technologies (IJCSIT)*, *5*(5), 6487-6490.
15. Wang, H., & Yoon, S. W. (2015). Breast cancer prediction using data mining method. In *IIE Annual Conference. Proceedings* (p. 818). Institute of Industrial and Systems Engineers (IISE).
16. Bahrami, B., & Shirvani, M. H. (2015). Prediction and diagnosis of heart disease by data mining techniques. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, *2*(2), 164-168.
17. Zand, H. K. K. (2015). A comparative survey on data mining techniques for breast cancer diagnosis and prediction. *Indian Journal of Fundamental and Applied Life Sciences*, *5*(s1), 4330-9.
18. Shukla, S., Gupta, D. L., & Prasad, B. R. (2016). Comparative study of recent trends on cancer disease prediction using data mining techniques. *International Journal of Database Theory and Application*, *9*(9), 107-118.
19. Assari, R., Azimi, P., & Taghva, M. R. (2017). Heart disease diagnosis using data mining techniques. *Int. J. Econ. Manag. Sci*, *6*(3), 1-5.
20. Almarabeh, H., & Amer, E. (2017). A study of data mining techniques accuracy for healthcare. *International Journal of Computer Applications*, *168*(3), 12-17.
21. Tafish, M. H., & El-Halees, A. M. (2018, October). Breast cancer severity degree predication using data mining techniques in the gaza strip. In *2018 International Conference on Promising Electronic Technologies (ICPET)* (pp. 124-128). IEEE.
22. Kaur, S., & Bawa, R. K. (2018, August). Review on data mining techniques in healthcare sector. In *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on* (pp. 224-228). IEEE.
23. Mia, M. R., Hossain, S. A., Chhoton, A. C., & Chakraborty, N. R. (2018, February). A comprehensive study of data mining techniques in health-care, medical, and bioinformatics. In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.
24. Eltalhi, S., & Kutrani, H. (2019). Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review. *IOSR Journal of Dental and Medical Sciences*, *18*(4), 85-94.
25. Mabu, A. M., Prasad, R., & Yadav, R. (2020). Mining gene expression data using data mining techniques: A critical review. *Journal of Information and Optimization Sciences*, *41*(3), 723-742.

26. Biwalkar, A., Gupta, R., & Dharadhar, S. (2021, May). An Empirical Study of Data Mining Techniques in the Healthcare Sector. In *2021 2nd International Conference for Emerging Technology (INCET)* (pp. 1-8). IEEE.
27. Arivazhagan, N., Mukunthan, M. A., Sundaranarayana, D., Shankar, A., Vinoth Kumar, S., Kesavan, R., ... & Abebe, T. G. (2022). Analysis of skin cancer and patient healthcare using data mining techniques. *Computational Intelligence and Neuroscience*, *2022*(1), 2250275.
28. Patela, R. (2023). Diagnosis of Breast Cancer using Data Mining Techniques.