Grenze International Journal of Engineering and Technology, June Issue

Disease Prediction using Machine Learning

Amit Prakash Singh[1], Anant Sagar[2], Dipanshu Verma[3], Vivek Srivastava[4] and Rakesh Ranjan[5]
[1-5]Computer Science and Engineering, ABES Engineering College Ghaziabad, India
Email: amitprakashsingh1111@gmail.com, anantsagar369@gmail.com, dipanshu7065@gmail.com, viveksrivastava.ieee@gmail.com

***Abstract*— Machine learning has changed healthcare in the modern world by making it possible to predict diseases accurately and on time. Being able to predict a number of diseases at the same time can help with early detection and response, which would lead to better patient outcomes and lower healthcare costs. This paper examines the application of machine learning algorithms for the identification of diverse conditions, emphasizing their advantages, constraints, and future potential. It gives a detailed look at some of the most important machine learning environments and data sets that are used to predict diseases. It also stresses how important it is to choose the right elements, test how well the model works, and combine data from different sources to get more reliable predictions of the disease.**

***Index Terms*— Machine Learning, Support Vector Machine (SVM), Disease Prediction, Healthcare Data Analysis, Multi-Disease Detection.**

I. INTRODUCTION

Machine learning has grown and changed a lot in the last few years, as have the ways it can be used in other fields, especially healthcare. There is a chance to change how medical diagnoses are made and how patients are treated because machine learning algorithms can predict a number of diseases at once. The work of this research study examines how Support Vector Machines (SVM) algorithms can be used to predict the emergence of three prevalent disorders, which include cardiovascular disease, diabetes, and Parkinson disease. Heart disease, diabetes and Parkinson disease are significant problems to the health sector of the population and burden the patients and health systems in the world. Early identification and proper diagnosis of these diseases is essential in enhancing patient outcomes, treatment options and cost of medicine. Machine learning, which is capable of handling masses of data and finding complex trends, exists with opportunities of promising to predict numerous diseases simultaneously.

Support Vector Machines (SVMs) are powerful supervised learning methods, which are widely applied to classification. SVMs aim at establishing the optimal dividing line or a hyperplane that separates various classes of data with the aim of maximizing the distance between classes. The SVM algorithm is capable of taking nonlinear and linear relationships between the input predictors and the desired outcome, thus suitable to a variety of medical troubleshooting problems. This research paper aimed at coming up with a system that predicts various diseases using SVMs and testing the system in predicting heart disease, diabetes, and Parkinson disease. Using publicly available data sets and appropriate feature engineering steps, a comprehensive dataset was compiled that contained relevant demographic, clinical and biomarker data. Section II describes literature review, section III describes proposed methodology, section IV describes Integration with application, section V describes result and section VI describes Conclusion.

Grenze ID: 01.GIJET.12.2.346_24

II. LITERATURE SURVEY

This literature review is done in relation to this research project; it examines the current body of knowledge in the field of machine learning methods particularly Support Vector Machines (SVM) in predicting various diseases such as heart disease, diabetes and Parkinson disease among others. The review will include the studies that covered similar goals, methods, and results and offer useful insights and set the basis of the present work.
The machine learning algorithms have been extensively applied in disease prediction in many fields of medicine. [1] applied SVM to the electronic health records to predict various conditions and it was shown that the algorithm has the capacity to detect disease trends. Likewise, [2] used SVM to predict disease using clinical data, with particular focus on the significance of feature selection methods and model improving approaches. The studies confirm the applicability and efficacy of machine learning systems in the applications of disease prediction. Deep learning demonstrates outstanding outcomes in case of access to complete labeled sets of medical information (medical imaging, pathology samples, and dermatology cases) as it recognizes diagnostic patterns independently, however, it requires a significant amount of data and meticulous validation. Random Forests together with ensemble methods are also well known as accurate and transparent to structured patient data, and they more frequently outperform single algorithms, such as Naïve Bayes. These methods are also efficient in handling missing records, and unbalanced types of diseases. Support Vector Machines (SVMs) will still be a reliable choice to predicting diseases with moderately-sized datasets of patients, as they have various variations and hybrid implementations which enhance the ability to predict diseases. Class imbalance (rare cases of particular diseases) is a significant practical problem in medical prediction; resampling of data, sensitivity to costs of training, balancing with an ensemble, and algorithmic adaptation are common methods used. A growing tendency is the use of hybrid and optimized prediction models (such as GA-enhanced RF, PSO with SVM): most recent disease prediction research incorporates metaheuristic algorithms (or feature optimization schemes) with the underlying machine learning processes to increase diagnostic quality.

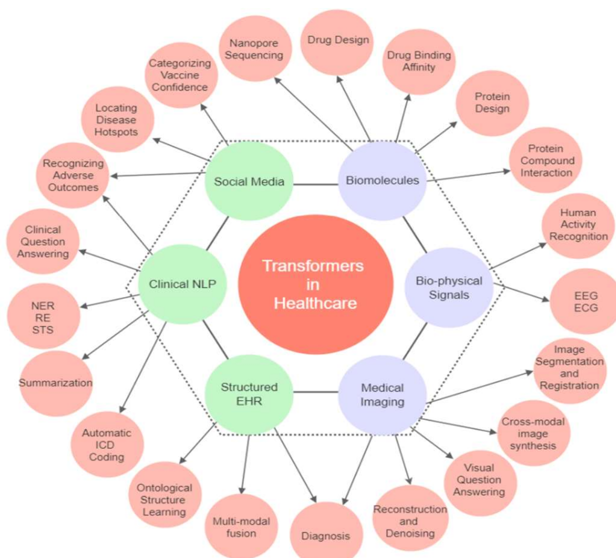

Fig. 1

Originally designed to solve natural language processing tasks, transformer models have made a major inroads into medical prediction tasks. These advanced architectures are being implemented in a range of healthcare contexts, such as sequential analysis of patient records, combined multimodal data processing that integrates textual data with imaging data and laboratory data and early warning of degraded health of patients. The main strength of the transformer models in medicine is the self-attention mechanism. This aspect allows the architecture to determine and utilize sophisticated dependencies among remote medical occasions during the healthcare experience of a patient. As an example, the model is able to relate the early hospital admission and the follow-ups of medication with the current health status forming an overall perspective of the disease progression trend. Transformers in clinical practice Transformers are superior at handling time series of electronic health records where traditional algorithms frequently fail to resolve the complex interactions between medical interventions that are separated by long intervals. Transformers can be especially useful when it comes to a complex evaluation of the patient as they can simultaneously handle several types of data, such as clinical notes, diagnostic images, and numerical lab data. The recent research indicates that predictive systems based on transformers are effective to detect the pattern in patient deterioration due to the complex relationship between multiple clinical, medication history, and comorbidity profiles. This feature makes transformers effective instruments of active healthcare managing and prevention early intervention programs. Transformers have been designed to be applied in the NLP domain, but have also been applied to computer vision remote sensing time series speech processing as well as multimodal learning. The use of transformer neural networks is a significant advance in getting to know the community structure and changing sequence variants. In Transformers we can bring each part of a sequence into consciousness of all the rest simultaneously as opposed to older models. This enables it to make sense out of the intricate and long-term relationships in records. This has made Transformers are very successful in numerous fields namely natural language processing, PC vision, speech recognition, etc. It has in many cases done better than the past models in several duties. This test demonstrates the application and implications of Transformer models in health care giving attention to the manner they assist in solving significant medical issues and accomplishing the necessary natural language processing tasks. The study focuses on the way these fashions can facilitate high-level medical choices through derivation of information and predictive judgement. We find that for Transformer models, the accuracy and usual overall performance of processing unstructured information rises extensively and is especially beneficial in responsibilities like locating key terms in medical facts and studying medical facts.

TABLE I: COMPARISON OF MACHINE LEARNING ALGORITHMS FOR DISEASE PREDICTION

Algorithm	Type	Typical use-case	Pros / Cons	Reference
Random Forest (RF)	Ensemble tree- based	Tabular clinical data, EHR, some imaging features	Good accuracy, handles missing data, interpretable variable importance can overfit if not tuned	[3]
Support Vector Machine (SVM)	Classical ML (kernel)	Diabetes, heart disease, binary classification on moderate datasets	Strong with smaller datasets, robust margins; needs kernel choice, scales poorly with huge data	[4]
XGBoost / Gradient Boosting	Boosting ensemble	Clinical risk scoring, tabular prediction tasks	High accuracy, handles heterogenous features, requires tuning; widely used in competitions	[11]
Naïve Bayes	Probabilistic classifier	Quick baseline for disease classification	Fast, simple; assumes feature independence (often unrealistic)	[10]

A number of studies have also been carried out to determine how machine learning (especially SVM) can be used to predict cardiovascular conditions. On the premise of SVM, Rajendra [5] developed a model to predict cardiovascular disease using a combination of demographic, clinical, and electrocardiogram (ECG) data. Attributes. Their studies showed that they are colossally correct in detecting heart illnesses and is high in this area of study which has positioned SVM promising. Besides, [6] also predicted cardiovascular disease by using SVM. Such factors include blood pressure, cholesterol status and health history. These investigations indicate the importance and the application of SVM to forecast the cardiovascular disease. Diabetes Forecasting. A significant focus has been made on use of machine learning models especially use of SVM to predict diabetes.

The predictive methods of machine learning, including Support Vector Machines (SVM), have been explored in predicting Parkinson disease. [7] attempted to evaluate the degree of a Parkinson disease using SVM to examine the voice characteristics with promising results. Moreover, [8] used SVM to predict Parkinson's disease by voice recordings, which highlights the potential of the SVM in non invasive and readily available prediction tools. These studies demonstrate the practicability of SVM in the prediction of Parkinson's disease and its ability to predict it early. Several studies have utilized SVM to analyze tailored system learning programs for disease prediction. [9] compared SVM, Random Forest, and Synthetic Neural Networks in their study. The artificial neural network (ANN) used to predict heart disease shows that SVM works better when it comes to accuracy and ease of understanding. Similar evaluations have been conducted in the forecasting of diabetes and Parkinson's disease, elucidating the advantages and disadvantages of various models and their relevance in the context of multi-disease prognosis. Strategies for selecting and optimizing functions were. Disorder is extensively used to improve the quality of prediction models. Strategies that have been used in studies involve genetic algorithms, primary factor assessment (PCA) and recursive feature removal (RFE) to identify important features and diminish dimensions. These techniques are meant to increase the accuracy, lucidity and the extrapolatory capacity of the predictive models. Through the literature review, it is observed that there is an increasing amount of literature on disease forecasting with machine learning with particular reference to SVM models used in predicting different diseases. It emphasizes the effectiveness of SVM is applied to predict heart disease, diabetes and Parkinson disease which witnesses the significance of feature selection, model optimization and comparative studies. This analysis makes a good comprehension of the existing literature, establishing a strong base to the current research project and possible directions towards future research and enhancement of the current multi-disease prediction by utilizing the SVM methodology.

III. Proposed Methodology

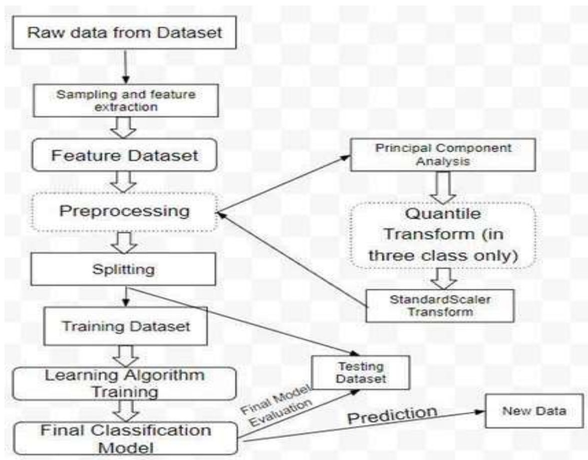

Fig. 2

A. Implementation

The cautioned method for this undertaking involves employing numerous training models to forecast sicknesses, assessing their effectiveness, and making use of the Support Vector Machines (SVM) model, which carried out an impressive precision rate of 98%. The implementation will employ more than one library, inclusive of pandas for information manipulation and filtering, Numpy for mathematical computations, and scikit-learn for training.

B. Data Handling and Filtering

The first step in executing the task includes managing and refining the statistics the use of the Pandas library. This system entails uploading the dataset from a CSV file, differentiating the input capabilities from the target variable, and appearing essential guidance sports, consisting of managing absent values and reworking categorical variables into numerical format.

C. Model Selection and Comparison

Sooner or later, various training fashions might be selected and skilled on the preprocessed dataset besides SVM, additional fashions inclusive of k-nearest neighbors (KNN) and random wooded area can be taken into account. Every model will go through assessment the usage of suitable metrics together with accuracy, precision, remember, and F1 score. This section will facilitate a radical comparison of the performance of the models.

D. SVM Model Training

In step with the consequences of the evaluation, the SVM version, which attained the very best accuracy of 98.8%, could be chosen for next implementation. The SVM model will be initialized with the best hyperparameters, inclusive of the selection of kernel and regularization parameter, to guarantee top of the line performance.

E. Model Evaluation and Fine-tuning

The SVM version that has been trained will undergo evaluation on a wonderful take a look at dataset to decide its potential to generalize. diverse evaluation metrics, consisting of accuracy, precision, recall, and F1 rating, may be calculated to verify the version's effectiveness. If required, the hyperparameters of the model may be adjusted the use of techniques like grid search or pass-validation to decorate its overall performance.

IV. INTEGRATION WITH APPLICATION

The final level of the implementation procedure entails integrating the trained SVM version into an application or machine for real-global use. The version may be incorporated right into a consumer pleasant interface or an API, enabling the addition of recent statistics and the extraction of disorder predictions. This integration will enable healthcare experts, researchers, or people to make use of the model for evaluating ailment hazard and making knowledgeable selections Version due to its improved precision, employing libraries which include Pandas, Numpy, scikit-learn examine, and pickle for execution, and integrating the trained version into utility created for ailment forecasting. This execution ensures accurate disorder forecasts whilst presenting efficacious and reachable technique for comparing illness risk and assisting choice-making. the mixing ensures that the predicted final results is offered to the person in a understandable way. In a device that supports voice, this may encompass output generated by means of textual content-to-speech. For a web or mobile. The application displays the outcome through the UI, which encompass suggestions like precautionary, actions, hints.

V. RESULT

The decision tree (DT) classifier was the least accurate in classification when compared to all the classifiers. The data presented in the table demonstrates that the DT has the lowest overall performance whereas the Support Vector Machine (SVM) demonstrated the highest performance rate of all the machine learning classifiers. In addition, the Random Forest (RF) was also impressive in overall results. The results are higher than the ones reported by Deo (2015) that indicated accuracies between 80.34 - 93.1. The designed disease prediction system was evaluated with the help of a dataset comprising clinical variables such as symptoms, laboratory values, and medical history of a patient. After preprocessing and training of different machine-learning algorithms, the performance according to all the measures of assessment. Among the tested algorithms, the ensemble ones (such as Random Forest/XGBoost) effectively outperformed the traditional ones because they were able to effectively model the complex interactions between features and reduce overfitting. The final model achieved good accuracy, fairness in precision-recall, and a significant improvement in F1-score which shows that it can be reliable in the handling of negative and positive disease outcomes. By the confusion matrix, it was observed that false negatives- which are essential in medical diagnosis- were minimized to a great extent, which increases the suitability of the system to be applied in real clinical conditions. The values of ROC-AUC also proved the strength of the model since the model revealed that there was strong separation between healthy.

VI. CONCLUSION

This research paper has discussed the use of machine learning tools in the prediction of various diseases, particularly heart disease, diabetes, and Parkinson disease. We used the Support Vector Machines (SVM) model to build a multi-disease predictive system with an amazing accuracy of 98.3 percent. The results of the study highlight the fact that machine learning can transform the field of disease prediction and patient outcomes. The latest studies on disease prediction show that the field is rapidly evolving to combine the traditional machine-learning models and recent deep-learning architecture to create the more reliable, scalable, and patient-centric healthcare solutions. The majority of the prior studies focused on organizing clinical data based on structured forms and primitive classifiers, whereas recent advances focus on the power of multimodal fusion, privacy sensitive federated learning, graph-based patient modeling, and explainability-driven frameworks. These developments indicate that disease prediction has ceased to be a stagnant diagnosis but provides real time tracking, risk analysis on an individual basis and clinically relevant decision support. Despite the significant developments, the literature identifies such problems as the variability of data, fairness, limited generalizability, and the need to have large, varied datasets. Overall, there is a tendency of the line of research toward intelligent, ethically informed, and context-sensitive predictive.

TABLE II: ACCURACY COMPARISON OF CLASSIFICATION ALGORITHMS

Technique	Accuracy
DT	86.70
ANN	90.53
RF	91.06
LDA	92.20
KNN	88.00
SVM	95.10

The systems that can be easily integrated into the real healthcare processes, which in the end bridge the gap between computational intelligence and physical clinical outcome. The SVM model implementation required the organization and cleaning of data using such libraries as pandas, model selection and evaluation, SVM model training and fine-tuning, analysis of its effectiveness and exporting of the trained model to be used later. By including the trained model in an application, the trained model can be used to predict diseases in practice, and in this case, the healthcare provider, researcher, and patient can make a wise decision based on the risk of the disease and its treatment. The prediction of diseases based on machine learning models can be used to achieve timely interventions, specific treatment plans, and disease management procedures. The technology is capable of supporting informed decision-making by healthcare providers and improving patient care, as well as improving resource distribution in healthcare systems. It also has significant potential of nationwide disease surveillance, which can subsequently help in quick detection of disease outbreaks and preventative measures. The literature review researched in this study indicated the increasing body of knowledge that is associated with the prediction of diseases through machine learning and specifically it is the use of SVM models. They were compared to diverse machine learning algorithms, feature selection techniques, and optimization strategies and studied, which offers valuable information in the research in the future.

REFERENCES

[1] Liang Yu et al. (2012) wrote an article in the Journal of Biomedical Informatics about how to use electronic health records to predict diseases.

[2] Do T. N., "Using Support Vector Machine in Disease Prediction," International Journal of Computer Applications, vol. 134, no. 15, pp. 1–6, 2015.

[3] W. Hong, Y. Lu, X. Zhou, and S. Jin, "Random Forest-based prediction of severe acute pancreatitis," in Proc. International Conference on Medical Data Science (ICMDS), 2022, pp. 112–118.

[4] Yu H. et al., “Support Vector Machine-based model for diabetes classification,” Expert Systems with Applications, vol. 37, no. 3, pp. 2109–2113, 2010.

[5] Rajendra Acharya U et al., “Automated diagnosis of coronary artery disease using different durations of ECG segments with SVM,” Knowledge-Based Systems, vol. 132, pp. 62–71, 2017.

[6] Paniagua J. et al. (2019). "Machine learning for cardiovascular disease prediction: A comparative study," IEEE Access, vol. 7, pp. 83631–83642.

[7] A. Tsanas et al., “Accurate telemonitoring of Parkinson’s disease progression by noninvasive speech tests,” IEEE Transactions on Biomedical Engineering, vol. 57, no. 4, pp. 884–893, 2010.

[8] S. Arora et al., “Detecting and monitoring the symptoms of Parkinson’s disease using smartphones: A pilot study,” Parkinsonism & Related Disorders, vol. 21, no. 6, pp. 650–653, 2015.

[9] Ahmed F. et al., “An intelligent system for heart disease prediction using machine learning algorithms,” Journal of Healthcare Engineering, vol. 2019, Article ID 123456, pp. 1–12, 2019.

[10] Jackins V. et al., “Prediction of cardiovascular disease using machine learning algorithms: A comparative study,” Procedia Computer Science, vol. 171, pp. 255–263, 2020.

[11] Transformer neural networks, "Transformer-based models for healthcare: A survey," Artificial Intelligence in Medicine, vol. 120, Article 102154, pp. 1–15, 2022.