Analysis of SVM-based Machine Learning Techniques to Improve the Accuracy of Disease Prediction in the Healthcare Sector

Siva Hari Naga Shashank Varagani
Independent Researcher, Staff Software Engineer Florida, USA.
Email: sivanagv12@gmail.com

Peer Review Information	Abstract
	This study employs an ensemble feature selection model—which combines wrapper and filter techniques—to extract critical characteristics from a dataset, with a focus on the UCI repository's benchmarking heart disease dataset. To assess the precision of the predictions, numerous machine learning approaches were fed the features that were ultimately chosen. Using a radial basis function, an enhanced Support Vector Machine (SVM) was suggested to improve healthcare prediction accuracy. In order to choose better support vectors, this SVM- Radial bias technique was fine-tuned using Particle Swarm Optimisation (PSO). Using the Chronic Kidney Disease (CKD) dataset from the UCI repository, the model's performance was tested extensively. The suggested model outperformed baseline machine learning algorithms like Random Forest (RF), Decision Trees (DT), and conventional support vector machine (SVM) techniques, with a remarkable accuracy of 98.7 percent.

Introduction

Problems with healthcare data overflow, a lack of medical professionals, and understaffed hospitals are problems that hospitals around the world have been experiencing. So, it's clear that any technology that can provide quick and effective data analysis for treatment plan development and workflow improvement in hospitals is highly valuable [1]. Within this framework, machine learning has emerged as a valuable resource for healthcare organisations seeking to collect and organise patient data, spot healthcare patterns, propose treatments, and much more besides.

What is machine learning in healthcare?

Machine learning is an effective method for teaching computers new skills by observing and mimicking human behaviour. In the absence of human operators, ML algorithms are able to learn from data on their own using mathematical techniques. Machine learning algorithms work better when there is more data to learn from. To rephrase, algorithms are able to learn from data patterns, draw inferences, and apply what they've learnt. Machine learning algorithms can analyse and handle massive volumes of healthcare data, revealing insights that would be hard to discover by hand [2]. Not only that, but they can assess risks, forecast outcomes, and do a whole lot more besides. Contrary to popular belief, ML-based solutions are not intended to supplant human healthcare providers. These instruments are made to complement and help them. Medical therapy and drug discovery are two areas where ML solutions are making a positive impact.

How does machine learning work?

There are various methods that can be employed to train ML models. They change according on the data sets given and the level of human intervention in their pre-processing. There are two types of data sets: raw (unlabelled) and labelled (data scientists-prepared) [3]. Knowing what data will be

available, if data scientists can be hired to prepare data sets, and the precise problems that need solving are all crucial for choosing the correct approach. When training ML models, three primary methods are employed:

- Supervised learning,
- Unsupervised learning,
- Semi-supervised learning.

Supervised learning

In addition to collecting data, there are a lot of pre-processing processes needed for the supervised learning method to use collected data. From this perspective, ML algorithms can accomplish the following tasks by combining input data with known responses:

- Master the art of learning input-output patterns for use in making predictions using fresh data.
- Locate things and attributes that allow for their categorisation.

Supervised learning is thus mostly employed for problems involving categorisation and predictions [4-8]. Supervised learning is often believed to be accurate; however, its accuracy is susceptible to the following factors:

- To get erroneous classification results, input features should not be picked carelessly.
- Prediction accuracy is impacted by training values that are incomplete or highly unlikely.

Furthermore, in order to create massive quantities of high-quality data sets, this method necessitates substantial human and computational resources. Supervised learning is still extensively utilised in healthcare, even though it has these disadvantages. Several supervised learning algorithms and the tasks they can do are illustrated below:

- A decision assistance tool for categorising various tumour types can be Decision Trees.
- Applications of Support Vector Machine include cancer diagnosis and handwriting recognition, among others.
- Naïve Bayes is applicable to prediction of radiation results and recommendation systems.

Unsupervised learning

Machine learning algorithms employ unlabelled datasets in unsupervised learning. Without any guidance or human input, these algorithms are able to find patterns or information in unstructured data. In order to do unsupervised learning, it is necessary to gather and observe data without labels.

Unsupervised learning algorithms are able to locate features to utilise for classification and find practically any pattern in data because they do not rely on human supervision. This allows them to do more complex jobs than supervised learning algorithms. Unsupervised learning methods often detect anomalies, reduce dimensionality, and cluster data [9, 10]. Genomic sequence analysis, objective recognition, market research, and the accurate selection of features (variables) for healthcare data pre-processing are some of the domains where unsupervised learning algorithms are used in healthcare.

Be wary of jumping to conclusions when using this method; unsupervised learning yields findings that aren't always predictable and require cautious interpretation. In order to help you choose the right technique, we've laid down the key differences between them below. Supervised and unsupervised learning differ in several ways, one of which is the use of labelled data.

Goals

Predicting results from new data with a pre-defined result type is the goal of supervised learning. Obtaining insights from fresh data is the objective of unsupervised learning. Hidden patterns in data sets can be defined by unsupervised learning techniques themselves.

Applications

A task that requires forecasting or predicting is best tackled via supervised learning. Anomaly detection, recommendation making, and medical image interpretation are all areas that benefit from unsupervised learning.

Complexity

This method is rather easy in comparison to unsupervised learning, even if supervised learning does necessitate preparation. Unsupervised learning models necessitate robust computational resources due to the substantial volumes of unclassified data they suggest. Furthermore, additional data is required for unsupervised learning models to provide results.

Literature Review

The importance of ML in meeting and analysing large volumes of data is highlighted in [11]. As a result, organisations can use this data to develop strategies that give them a competitive edge. By analysing this data critically, organisations can make better decisions quickly, which strengthens their position in the market. In addition to helping with cost reduction, advanced ML and other technologies allow organisations to do things like speed up product delivery by lowering transportation and inventory costs. This is just one example of how management may use these tools to their

advantage.

In order to better engage with customers, understand their needs, and provide value-added products and services, businesses are increasingly utilising machine learning and deep learning technologies. Additionally, ML helps the company learn more about its clients, examine the market's geographical division, and provide services that boost satisfaction. A.I. is making its way out of the research and development lab and into the commercial sector. All around the globe, millions of top sectors and businesses are harnessing the potential of AI and AAI [12]. When it comes to detecting fraud and improving consumer happiness, most sectors rely on millisecond machine learning algorithms. Machine learning, business platforms, and application-based solutions have proliferated in recent years, all with the intention of satisfying businesses. Not only can these cutting-edge technologies streamline the software and Internet businesses, but they also improve other sectors including construction, healthcare, law, automobiles, and agriculture.

The majority of major corporations invest in AI R&D for the benefit of their consumers and other businesses. This includes IBM, Facebook, Apple, Google, Amazon, Myntra, Flipkart, and Microsoft. Machine learning algorithms and artificial intelligence can use digital recordings of human behaviour to forecast some private functions and features. We will be more efficient and capable of handling more duties with the support of AI and a deep learning model [13]. Support for voice recognition, Google, Alexa, and Siri are the most obvious manifestations of ubiquitous artificial intelligence. There are massive gatherings. Evidently, these platforms target the workplace, and nearly half of all Europeans, Chinese, Americans, and the majority of Indians are on board with using speech recognition in some way. A number of companies have already incorporated Alexa into their business models to enhance company culture and productivity, including Brooks Brothers, WeWork, Mitsui USA, Capital One, and Vonage. If you possess a state system, you can't see its progress or ownership. When we examine these boundaries, new ideas emerge. Machine learning and artificial intelligence-enabled creatures to overcome these grey zones. It may involve evaluating the patient's risk or introducing a new analytical tool in the medical field. This entails anticipating mistakes in the building process before they happen [14]. Internet and intranet-connected devices are the talk of the town these days. Businesses may now gather more digital data, get the insights they need, and transform their thinking and operations. There will be much-needed market development, including quicker markets, easier regulations and processes, dynamic transactions, higher returns, informed customers, motivated transactions, and well-informed customers [15]. Despite the fact that most AI inventions lead to economic development through the recognition of novel openings and interruptions and are aesthetically pleasing in modern action, one market segment that other business moguls have overlooked or that is entering new markets offers connected devices [16]. Exciting financial breakthroughs are made even more so by a constant flow of data, which feeds machining, production, an automated support centre on demand, robots, and more. The way nations learn the fundamentals of computer science in areas like cyberspace, spacecraft, and aviation will have a profound impact on the AI-powered digital business revolution of the future. If we want to alter the AI landscape, we need to pay close attention to data-driven companies that can scale to collect more data, information, and expertise.

Problems

Supervised learning takes more effort and need more experts in data labelling than unsupervised learning.

Unsupervised learning produces results, but human operators must demonstrate their correctness and practicality.

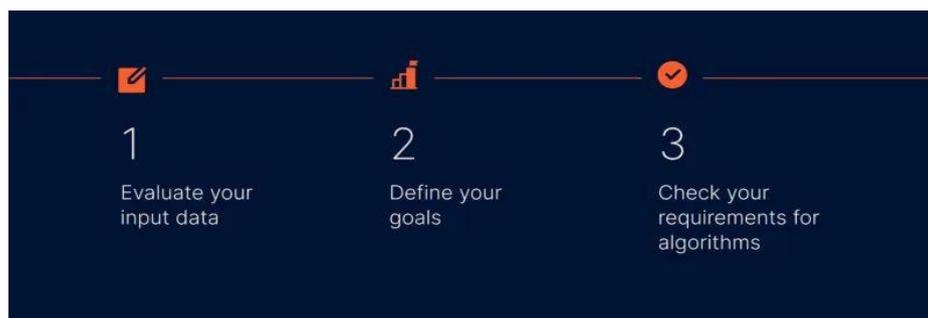We advise that you follow these steps to ensure that you make the proper decision are shown in figure 1:

Fig 1: Unsupervised learning problem

It is difficult to extract useful features from large datasets, which is why semi-supervised learning is useful in these situations. While a limited percentage of data can be labelled (e.g., 10%), semi-supervised learning datasets still comprise both unlabelled and labelled data. Medical imaging stands to benefit greatly from this method: by classifying MR pictures of malignant tumours exclusively, a computer can identify individuals who need further evaluation in this area.

Healthcare Use Cases For Machine Learning
For a better understanding of how ML technologies could improve healthcare, have a look at the examples provided below figure 2.
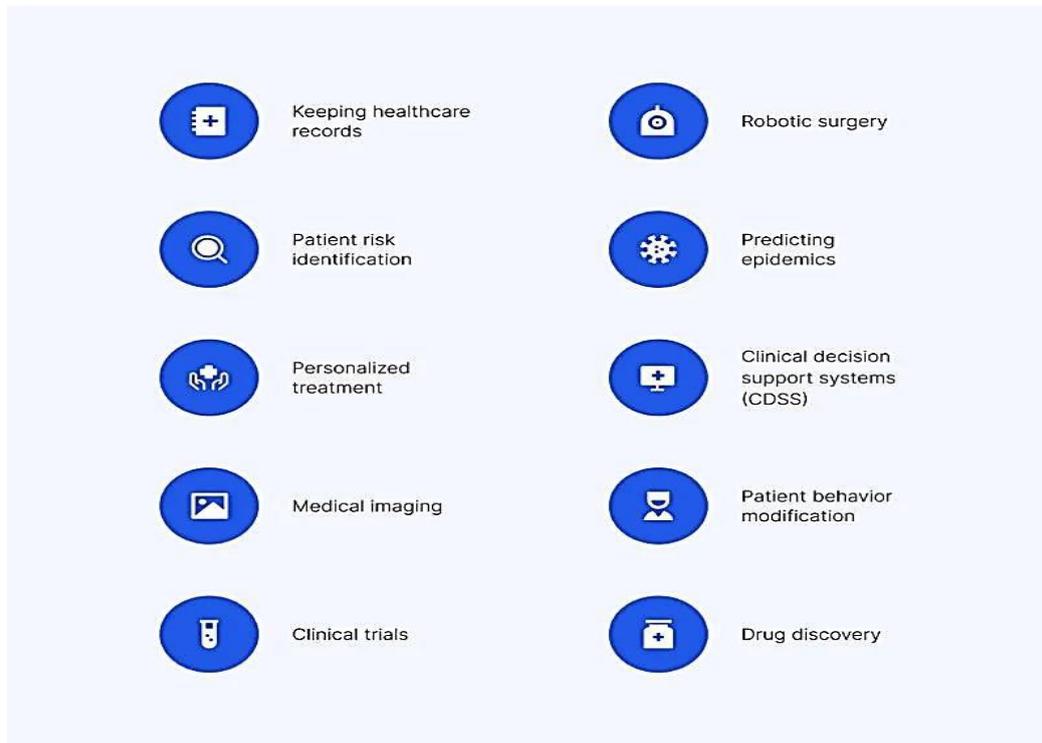

Fig 2: 10 examples of machine learning in healthcare

Keeping healthcare records
Tools based on ML can make it easier to maintain electronic health records (EHRs). This difficult assignment requires constant input of new data as well as revisions to current data. Medical record keeping is still laborious, error-prone, and plagued by broken interfaces and repetitive tasks, even with technological developments. Medical speech recognition, text recognition, and advanced analytics capabilities are all part of smart healthcare record systems powered by ML technologies. The digitisation of paper copies, the transcription of spoken conversations between physicians and patients, and the preservation of current and accurate patient health data are all made possible by these technologies. On top of that, these solutions can help doctors with diagnosing and making decisions.

Patient risk identification
One of the most exciting uses of ML in healthcare is patient risk prediction. The state of a patient and the risk of illness development in the future can be predicted by training ML models with massive volumes of previous patient data. This paves the way for healthcare practitioners to proactively identify patients who are at high risk and give them with tailored treatment.

Personalized treatment
Standardised therapy procedures are provided by treatment protocols for certain conditions. But it's not uncommon for patients to have varying reactions to the treatment plans. ML approaches let clinicians to discover therapy choices that are tailored to each patient by analysing input data such as lab test results, genetics, nutrition, and more. Fig 3 shows the Using machine learning applications in healthcare for high-quality diagnostics.
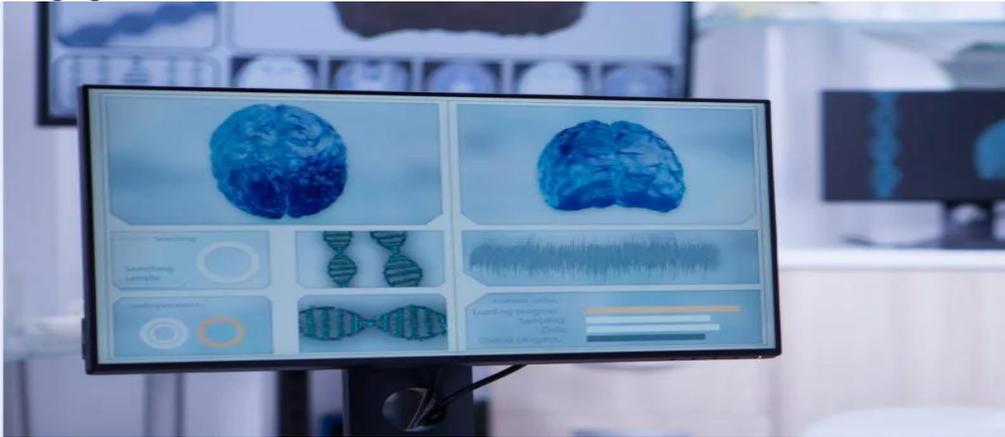
Medical imaging

Fig 3: Using machine learning applications in healthcare for high-quality diagnostics

Historically, radiologists and doctors have been responsible for interpreting medical pictures in healthcare. But there are a lot of diseases and changes that aren't always easy to see with the naked eye, so it might be tough. Machine learning algorithms can be taught to identify intricate patterns of cancer, even in challenging cases or in the early stages of the disease. Consequently, ML tools are proving to be quite useful for accurate diagnosis and effective treatment planning.

Clinical trials

There is a high potential for mistakes and deviations in the complicated, time-consuming, and costly procedures used in clinical research. If these things aren't taken into account, the trial's implementation and acceptance could end in failure. However, ML can optimise data collection techniques, analyse results, and help with participant selection, all of which increase the operational and scientific efficiency of clinical trials. Researchers may forecast clinical trial results, discover hidden trends, and efficiently manage big and diverse data sets by using ML- based solutions.

Robotic surgery

Robotic surgery is among the most significant areas where ML is being applied in healthcare and are shown in figure 4. The following are just a few examples of the many field-related activities that ML may handle:

Fig 4: Future of machine learning in healthcare

The process of suturing automated. Suturing is a necessary yet time-consuming aspect of many surgical procedures. Surgeons and patients alike would benefit from shorter operating times and less physical strain if the suturing process were automated. Procedure modelling in surgery. Modern surgical procedures necessitate constant refinement due to the increasing complexity of these procedures. Surgical process modelling is a novel strategy that attempts to improve the efficacy and security of modern complex operations. Surgical workflows can be modelled using ML approaches, with surgeon expertise and patient-specific information (such

as health state and clinical history) taken into consideration. So, ML can be quite useful for making fast and accurate surgical judgements.

Beyond actual surgery, models of surgical procedures can be utilised for tasks such as training and skill evaluation, workload analysis, technology introduction, and procedure length prediction.

Predicting epidemics

By analysing massive volumes of data, modern technologies can predict when epidemics will break out. Climate, demographics, geography, and other factors are considered in conjunction with the rate of epidemic transmission in order to make predictions about the future. Particularly in places where medical treatment is scarce, this aids in pinpointing possible disease epidemic hotspots and implementing measures to contain them.

Clinical decision support systems (CDSS)

Care decision support systems (CDSS) can improve the effectiveness of care teams by analysing massive amounts of data, recommending future treatment options, identifying possible problems, and more. Limited patient data is usually used to make conclusions about diagnoses. In a clinical decision support system (CDSS), individualised evaluations and suggestions are made possible by matching patient data with a computerised clinical knowledge base. Recent developments in CDSS development aid doctors in making more educated diagnoses, treatment recommendations, and outcome evaluations.

Patient behavior modification

Things like smoking and not getting enough exercise can quickly set off or exacerbate the progression of chronic diseases. By following healthy behaviours, which necessitates supervision and consistency, these factors can be controlled or eliminated. The implementation of ML-based solutions can greatly assist patients in making behavioural adjustments and embracing healthier lifestyle choices. Machine learning algorithms take activity-related data input from patients (e.g., wearables), process it to find patterns, then develop individualised treatment programs while continuously tracking their progress.

Drug discovery

A lot of money and time has to go into finding a good medicine. Finding the optimal mix requires testing thousands of molecules and combinations thereof. Machine learning algorithms, when applied to healthcare, may, nevertheless, predict chemical features with remarkable speed and accuracy, greatly improving the efficiency with which prospective medications can be identified.

Benefits of machine learning in healthcare

Finally, these are the key benefits that healthcare organisations stand to gain from applying ML technologies which are shown in figure 5.

Fig 5: Advantages of machine learning in healthcare

Personalization

Medical ML's capacity to deliver individualised treatment choices, suggestions, healthy living strategies, etc. is one of its most impressive aspects. With the customer-centric approach gaining traction in healthcare, one of the main objectives is to provide a tailored experience to every patient. Machine learning is a powerful tool for this endeavour.

Innovations

Research and medication development are two areas that make excellent use of ML techniques. Drug development costs and time-to-market can be reduced for pharmaceutical businesses by expediting drug discovery techniques, such as improved sequencing methods and cutting-edge manufacturing workflows. Clinicians gain from this since it gives them more options on how to treat their patients. Furthermore, ML systems can analyse data from trial participants, decrease data-based errors, and make participant selection easier for clinical trials.

Streamlined processes

Using machine learning in healthcare is crucial for a number of reasons, the most important of which are to deal with interruptions, improve

workflow efficiency, and satisfy patient needs.

Precise clinical decisions

Through the rapid processing of massive healthcare data sets, machine learning enables physicians to expedite diagnostics and enhance decision-making with intelligent predictions and conclusions returned in a matter of minutes.
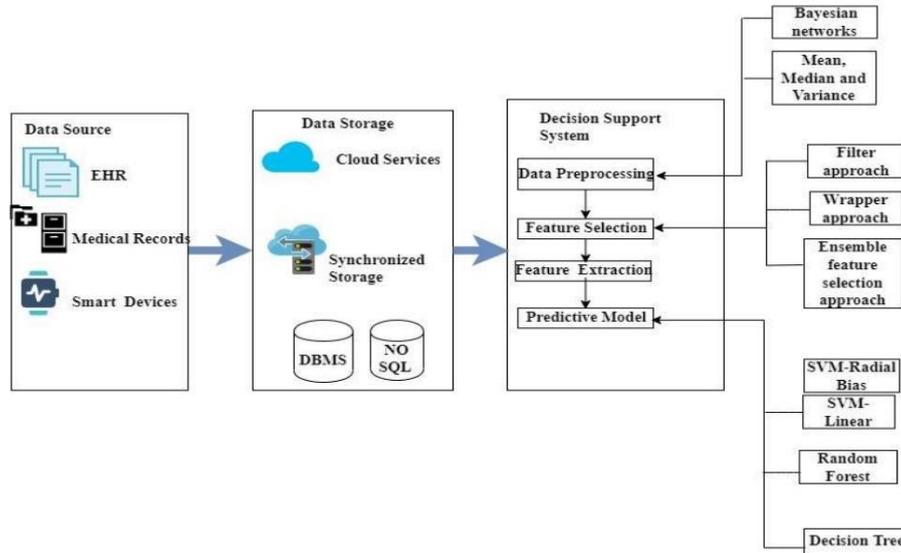
Methodology

Fig 6: machine learning-based decision support system in healthcare

Figure 6 described as:

1. Data Source:
   - **EHR (Electronic Health Records):** These are digital versions of patients' paper charts and are a rich source of patient health data, including medical history, diagnoses, medications, treatment plans, immunization dates, allergies, and lab test results.
   - **Medical Records:** This includes structured and unstructured medical data from various sources, such as doctor's notes, imaging reports, and lab test results.
   - **Smart Devices:** These include wearables or medical devices that continuously monitor and collect patient health data, such as heart rate, blood pressure, glucose levels, etc.
2. Data Storage:
   - **Cloud Services:** Cloud platforms where healthcare data is stored for scalability, remote access, and integration. These services provide high availability, security, and infrastructure for large-scale healthcare data management.
   - **Synchronized Storage:** Refers to systems ensuring that data stored across multiple platforms (cloud, local, etc.) is consistent and up to date.
   - **DBMS (Database Management Systems):** These are systems used to store structured data from healthcare records, often using SQL databases for querying.
   - **NoSQL:** For handling large-scale, unstructured healthcare data, such as patient notes, images, and continuous data streams from smart devices.
3. Decision Support System:
   - **Data Preprocessing:** Before data can be used for machine learning, it undergoes cleaning, normalization, transformation, and imputation of missing values to ensure that the data is usable for modeling.
   - **Feature Selection:** This involves identifying the most relevant features (e.g., patient characteristics, vitals, medical history) that will contribute to the predictive model.
   - **Feature Extraction:** Deriving new features from raw data through methods like dimensionality reduction, to create a simplified but informative dataset.
   - **Predictive Model:** The core machine learning model that makes predictions or recommendations. Different models and approaches are listed on the right.
4. Algorithms and Approaches shown in figure 7 is:
   - **Bayesian Networks:** These are probabilistic models that represent relationships among variables (e.g.,

symptoms, diagnosis) using conditional dependencies.

- **Mean, Median, and Variance:** These are basic statistical methods used in the feature selection process for understanding data distribution and variability.

- **Feature Selection Approaches:**
  o **Filter Approach:** Selects features based on their statistical relationship with the target variable (e.g., using correlation or chi-square tests).
  o **Wrapper Approach:** Uses machine learning models to evaluate the importance of different subsets of features.
  o **Ensemble Feature Selection Approach:** Combines multiple feature selection methods to improve the robustness and accuracy of feature selection.

- **Predictive Models:**
  o **SVM-Radial Bias:** Support Vector Machine (SVM) model using the Radial Basis Function (RBF) kernel for nonlinear classification.
  o **SVM-Linear:** SVM model using a linear kernel, typically used for linearly separable data.
  o **Random Forest:** A decision tree-based ensemble method that creates multiple trees during training and outputs the majority class for classification.
  o **Decision Tree:** A model that splits the data into branches based on feature values to make predictions, commonly used for both classification and regression tasks.

Algorithm: Training Model for SVM

Input: D=[X,Y]; X(array of input with m features), Y(array of class labels)
Y=array(C) // Class label
Output: Find the performance of the system
function train_svm(X,Y, number_of_runs)
initialize: lrate=Math.random();
for lrate **in** number_of_runs
error=0;
for i **in** X
if (Y[i] *(X[i]*w))<1 **then**
update : w=w + lrate * ((X[i]*Y[i])*(-2*(1/number_of_runs)*w)
else
update: w=w+lrate *(-2*(1/number_of_runs)*w)
end if
end
end

Fig 7: SVM Training Model.
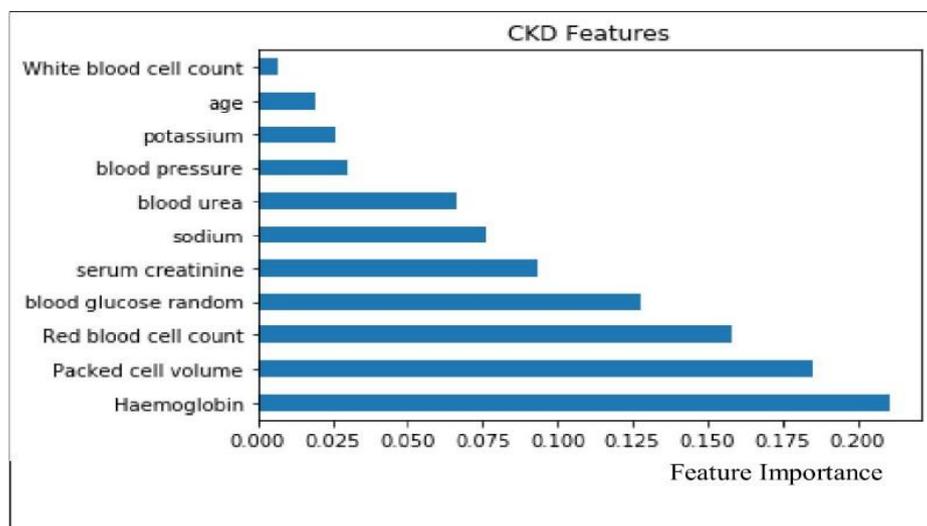
Results And Discussion

Fig 8: Important features in CKD Prediction using Chi-Square Score.

The chi-square score values of the features utilised in the model are displayed in Figure 8, which denotes the statistical significance of each feature in the dataset. As it reveals the association between characteristics and the dependent variable, the chi-square test is useful for feature selection.
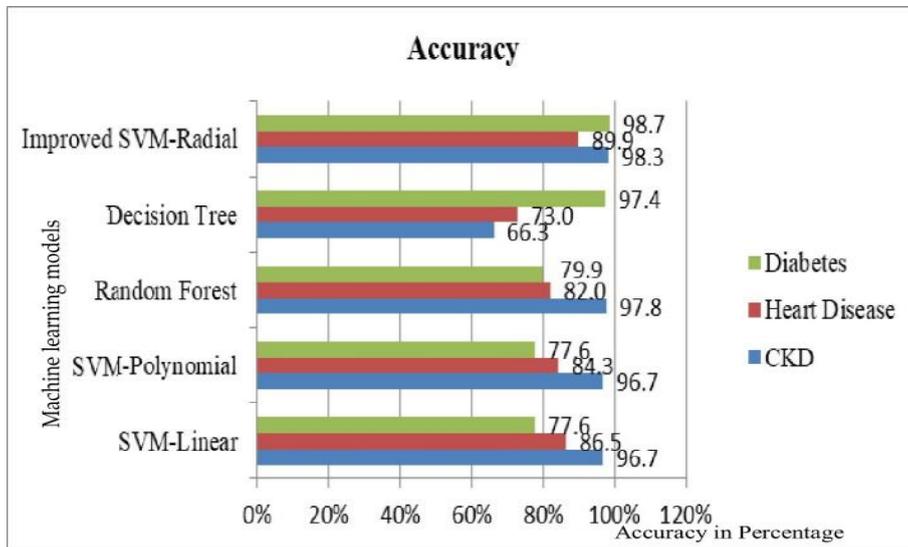

Fig 9: Performance Evaluation – Accuracy

The Performance Evaluation – Accuracy is shown in figure 9.
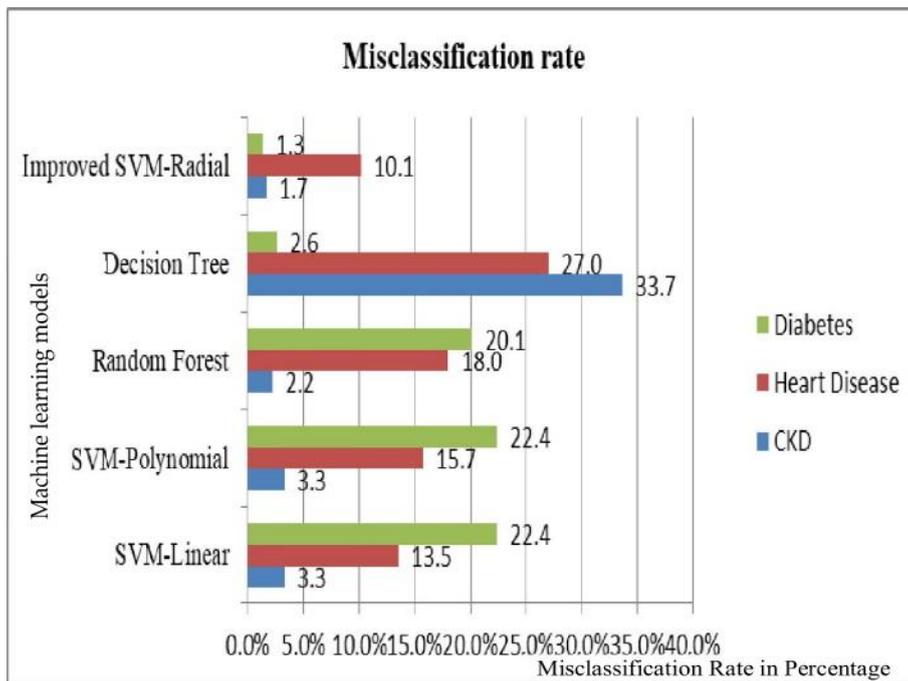

Fig 10: misclassification rate

Various machine learning methods and the suggested Improved SVM radial methodology are shown graphically in Figure 10 according to their misclassification rate. In the context of CKD illness diagnosis, the suggested optimisation method in the SVM reduces the misclassification rate from 3.3% to 1.7%. Similarly, the percentage of misclassification in diagnosing heart disease has been decreased from 13.5 percent to 11.1 percent. From 22.4% to 1.3%, the misclassification rate in diabetes prediction was significantly reduced.
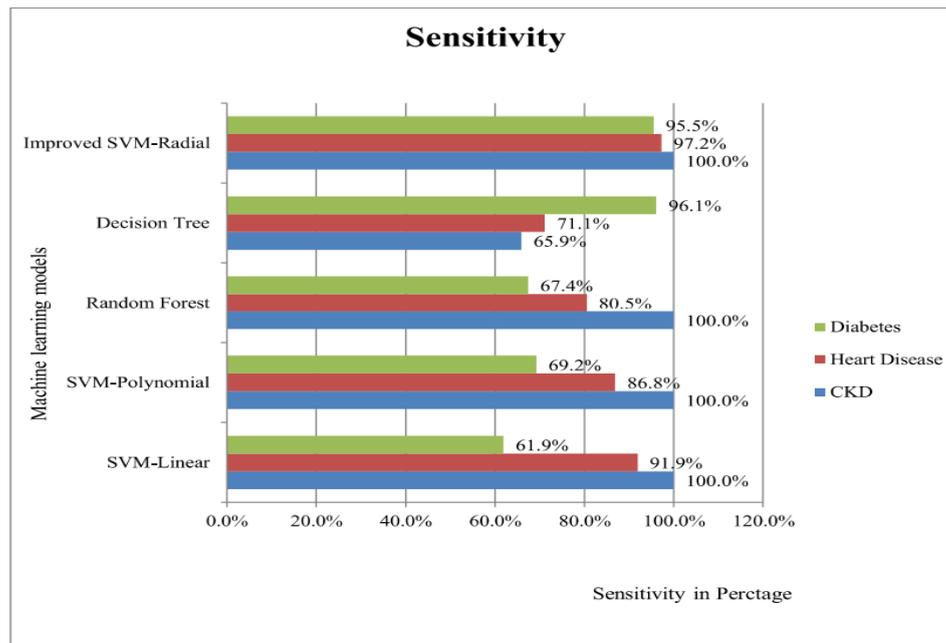
Fig 11: Sensitivity rate.

The Performance Evaluation of Sensitivity rate is shown in figure 11.

Conclusion

Proposed an ensemble feature selection model using the wrapper and filter approaches to determine the essential features in the dataset. The proposed model was implemented with a benchmarking heart disease dataset repository (UCI repository) to know the effectiveness of the model. Selected features from this proposed model were given as input for several machine learning techniques to analyze prediction accuracy. For the Decision Support System, an Improved SVM –Radial bias technique to improve the prediction accuracy in the healthcare industry was proposed. This SVM-Radial bias algorithm was improved by using particle swarm optimization to choose better support vectors in SVM. The model's effectiveness was tested in terms of accuracy by using the Chronic Kidney Disease (CKD) dataset from the UCI repository. The proposed model was compared with other standard machine learning algorithms such as Random Forest (RF), Decision Trees (DT) and other SVM algorithms. From the experimental results, it was found that the proposed model generated an accuracy of 98.7%.

References

Abiyev, RH & Abizade, S 2016, „Diagnosing Parkinson‟s Diseases Using Fuzzy Neural System‟, Computational and Mathematical Methods in Medicine, vol. 2016, pp. 1–9.

Aldhyani, THH, Alshebami, AS & Alzahrani, MY 2020, „Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms‟, Journal of Healthcare Engineering, vol.2020, pp. 1–16.

Alhaj, TA, Siraj, MM, Zainal, A, Elshoush, HT & Elhaj, F 2016, „Feature selection using information gain for improved structuralbased alert correlation‟, PLOS ONE, vol. 11, no. 11, pp. 1-18.

Altidor, W, Khoshgoftaar, TM, Van Hulse, J & Napolitano, A 2011, „Ensemble feature ranking methods for data intensive computing applications‟, Handbook of Data Intensive Computing, Available from: Springer, New York, NY[11 November 2011]. https://doi.org/10.1007/978-1-4614-1415-5_13, pp. 349–376.

Alyami, R, Alhajjaj, J, Alnajrani, B, Elaalami, I, Alqahtani, A, Aldhafferi, N, Owolabi, TO & Olatunji, SO 2017, „Investigating the effect of correlation based feature selection on breast cancer diagnosis using artificial neural network and support Vector Machines‟, International Conference on Informatics, Health & Technology ICIHT. https://doi.org/10.1109/iciht.2017.7899011

Amrieh, EA, Hamtini, T & Aljarah, I 2016, „Mining educational data to predict student‟s academic performance using Ensemble

Methods", International Journal of Database Theory and Application, vol. 9, no. 8, pp. 119–136.

Ana Clara Naufel de Felippe, Maria Helena Marotti Martelletti Grillo & Thaís Helena Grechi, 2006, „Standardization of acoustic measures for normal voice patterns ", Brazilian Journal of Otorhinolaryngology, vol. 72, no. 5, pp-659-664.

Anbarasi, M, Anupriya, E & Iyengar, NCSN 2010, „Enhanced prediction of heart disease with feature subset selection using genetic algorithm", International Journal of Engineering Science and Technology, vol. 2, no.10, pp. 5370-5376.

Anooj, P 2011, „Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules and decision tree rules", Open Computer Science, vol. 1, no. 4, pp. 482-498

Anooj, P 2012, „Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", Journal of King Saud University - Computer and Information Sciences, vol. 24, no. 1, pp. 27–40.

Battiti, R 1994, „Using mutual information for selecting features in supervised neural net learning", IEEE Transactions on Neural Networks, vol. 5, no. 4, pp. 537–550.

Behroozi, M & Sami, A 2016, „A Multiple-Classifier Framework for Parkinson"s Disease Detection Based on Various Vocal Tests", International Journal of Telemedicine and Applications, vol. 2016, pp. 1–9.

Ben Brahim, A & Limam, M 2017, „Ensemble feature selection for High Dimensional Data: A new method and a comparative study", Advances in Data Analysis and Classification, vol. 12, no. 4, pp. 937–952.

Benmalek, E, Elmhamdi, J & Jilbab, A 2017, „Multiclass classification of parkinson"s disease using different classifiers and LLBFS feature selection algorithm", International Journal of Speech Technology, vol. 20, no. 1, pp. 179–184.

Bikbov, B, Purcell, CA, Levey, AS, Smith, M, Abdoli, A, Abebe, M & Owolabi, MO 2020, Global, regional, and national burden of chronic kidney disease", 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet, vol. 395, no. 10225, pp. 709-733. 1

Bolón-Canedo, V & Alonso-Betanzos, A 2019, Ensembles for Feature Selection: A Review and Future Trends", Information Fusion, vol. 52, pp. 1–12.