IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 March 2024, accepted 7 May 2024, date of publication 13 May 2024, date of current version 30 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3400023

RESEARCH ARTICLE

Disseminating the Risk Factors With Enhancement in Precision Medicine Using Comparative Machine Learning Models for Healthcare Data

A. SHEIK ABDULLAH, (Member, IEEE), V. NAGA PRANAVA SHASHANK, AND D. ALTRIN LLOYD HUDSON

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu 600127, India

Corresponding author: A. Sheik Abdullah (sheikabdullah.a@vit.ac.in)

ABSTRACT The field of healthcare analytics, as an expanding discipline that integrates data analysis, machine learning, and clinical expertise, is seen to hold great promise for improving patient outcomes and the overall delivery of healthcare services. With the increasing availability of Electronic Health Records (EHRs), a wealth of healthcare data has emerged, presenting opportunities to enhance disease prediction and personalize treatments. The objective of the research is to develop and evaluate machine-learning models for predicting cancer, diabetic, diabetic retina, and heart-related outcomes using demographic and clinical data from Electronic Health Records (EHRs). Through thorough testing on diverse datasets, the study aims to assess the performance of these models in terms of accuracy, precision, and recall metrics, with the ultimate goal of advancing disease prediction and enhancing patient outcomes within the field of healthcare analytics. The proposed model demonstrates high accuracy, particularly in predicting cancer (97.080%) and diabetic (97.33%) outcomes using Support Vector Machines (SVM) and Decision Trees. Additionally, logistic regression achieves a notable accuracy of 76.521% for diabetic retina dataset, while Decision Trees exhibit 86.419% accuracy for heart-related predictions. SVM accuracy for Pima diabetic dataset stands at 79.746%. To assess the model's performance, thorough testing was conducted on a diverse and extensive dataset, employing a combination of accuracy, precision, and recall metrics. This research represents a substantial contribution to the field of healthcare analytics, emphasizing the potential of machine learning to advance disease prediction and, ultimately, enhance patient outcomes.

INDEX TERMS Machine learning, decision trees, logistic regression, random forest classifier, support vector machine, breast cancer, heart disease, PIMA Indian diabetes, diabetic retinopathy, diabetes, comparative analysis, hyperparameter tuning.

I. INTRODUCTION

Machine learning (ML) has undoubtedly revolutionized the healthcare industry, ushering in a transformative era that empowers the creation of sophisticated disease prediction models. These models are crafted to analyse extensive collections of medical data, providing valuable insights into the probability of a patient developing a particular disease or medical condition. ML algorithms, trained on a diet of electronic health records, medical imaging, genetic data, and a host of other sources, possess the remarkable ability to discern intricate patterns and features that serve as harbingers of particular diseases. With the knowledge distilled from this extensive training, ML algorithms can then make predictions when confronted with new patient data. This capability fundamentally revolutionizes disease detection, treatment, and even prevention in some cases. The appeal of disease prediction through machine learning lies

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

in its ability to streamline the challenging task of analysing medical data. By doing so, it not only reduces the workload on healthcare providers but also enhances the accuracy of predictions, leading to substantial improvements in patient care and results. The potential of ML-based disease prediction in healthcare is immense, offering opportunities for substantial enhancements in patient outcomes, reductions in healthcare expenses, and, most importantly, the preservation of lives. This research project distinguishes itself through its comprehensive comparative analysis of ML algorithms spanning a diverse spectrum of diseases. This encompasses afflictions such as cancer, Pima diabetes, diabetic retinopathy, heart disease, and more. By casting such a wide net, the research facilitates a holistic evaluation of algorithmic performance, shedding light on their adaptability and effectiveness in various clinical contexts. In addition to disease prediction, this study ventures into the territory of precision medicine. It does not stop at predicting diseases; it goes a step further by offering personalized treatment recommendations based on predictive models. This integration extends beyond traditional disease prediction, promising tailored care that can significantly improve patient well-being. However, the journey is not without its complexities. The study delves deep into the intricate ethical and privacy considerations associated with the implementation of predictive models in clinical practice. It recognizes the necessity of addressing these concerns to ensure the responsible and ethical use of these powerful tools.

One of the most exciting aspects of this research is the potential for precision medicine. Instead of providing standardized treatment recommendations for everyone, machine learning can assist in customizing treatments based on an individual's distinct genetic and medical characteristics. Customized strategies hold promise for yielding efficient therapies with minimized adverse reactions, ultimately elevating the overall health of patients. Precision medicine holds significant potential to revolutionize the management of conditions such as cancer [3]. By analyzing a patient's genetic makeup and the genetic characteristics of their tumor, oncologists can select the most appropriate and targeted therapies. This approach minimizes the use of ineffective treatments, reducing side effects and improving the chances of remission. This multidimensional approach, bolstered by the utilization of a vast, real-world dataset, underscores the originality and practical significance of this research endeavor. It positions this work at the forefront of healthcare decision-making and machine learning integration, guiding the industry toward a future where disease prediction and precision medicine become standard practice, ultimately transforming healthcare for the better. As we continue to advance in the field of machine learning and healthcare, it is essential to remember that these technologies are tools to assist healthcare professionals, not replace them. While ML can analyze vast amounts of data and make predictions, the human touch, empathy, and clinical expertise of healthcare providers remain irreplaceable. The collaboration between humans and machines can unlock the full potential of ML in healthcare, leading to better patient care, improved outcomes, and a healthier future for all.

Implementing machine learning in healthcare presents critical challenges, with data privacy standing out as a paramount concern. Electronic health records, medical images, and genetic data harbour sensitive patient information, necessitating robust measures like encryption and anonymization to safeguard privacy while enabling effective analysis. Achieving a delicate balance between data utility and privacy remains an ongoing challenge, requiring careful consideration to prevent overly aggressive anonymization that compromises data usefulness. Additionally, the potential for algorithm bias poses another complexity, as historical data may contain biases leading to disparities in healthcare outcomes. Addressing these challenges demands ongoing monitoring, bias mitigation strategies, and collaboration between regulatory bodies and the research community to establish guidelines ensuring the ethical deployment of machine learning in healthcare.

The partnership between machine learning and healthcare promises transformative advancements, offering opportunities to revolutionize medical practices. Disease prediction models and precision medicine advancements hold immense potential for improving patient outcomes. However, this journey towards advancement faces hurdles such as data privacy concerns, algorithmic biases, and regulatory intricacies. Overcoming these challenges is crucial to ensure the responsible and ethical integration of machine learning into healthcare. Stakeholders across the industry, including researchers, healthcare providers, regulators, and patients, must collaborate to navigate these complexities effectively and harness the full potential of machine learning for the betterment of healthcare. The ultimate goal is to establish personalized and effective care as the standard, with patient well-being at the forefront, ensuring a healthier and more equitable future for all.

As machine learning continues to shape the landscape of medicine, addressing key challenges becomes imperative for its successful integration into healthcare practices. Data privacy, algorithmic biases, and regulatory frameworks pose significant hurdles that must be overcome to realize the full potential of machine learning in improving patient outcomes. Achieving a balance between maximizing data usefulness while safeguarding privacy, addressing biases, and implementing rigorous validation and regulation processes are critical measures for ensuring the ethical and responsible application of machine learning in healthcare. It is important for stakeholders to collaborate closely to navigate these complexities and unlock the immense potential of machine learning to transform healthcare delivery.

II. RELATED WORK

Each year, the number of deaths attributable to breast cancer climbs considerably. Enhancements in the prediction and diagnosis of cancer are crucial for sustaining a

healthy lifestyle. Consequently, achieving a high level of precision in cancer prognosis becomes imperative for modernizing therapeutic approaches and elevating the quality of patient survivability. The provided references encompass diverse facets of detecting and predicting breast cancer through machine-learning techniques. The study conducted by Agarap [1] discusses breast cancer detection and is presented at the second International Conference on Machine Learning and Soft Computing in February 2018. The paper likely explores the use cases of ML algorithms in detection of breast cancer.

The study by Ganggayah et al. [2] centres on utilizing machine-learning techniques to predict factors influencing the survival of individuals with breast cancer. Published in BMC Medical Informatics and Decision Making in March 2019, the paper likely delves into the identification of key features and variables that influence the survival outcomes of individuals with breast cancer. This type of research is crucial for improving treatment strategies and personalized healthcare. In a study by Dalal et al. [3], a hybrid ML model for accurately predicting the breast cancer was introduced. The paper likely proposes a novel approach that combines various ML algorithms to enhance the accuracy and timeliness of breast cancer predictions. Hybrid models often leverage the strengths of multiple algorithms to address the limitations of individual methods. Moving beyond breast cancer, the study by Kimura [3], published in March 1986, discusses a structure editor for abstract document objects. Although not directly related to breast cancer, this reference could be relevant in the context of understanding the historical development of software engineering tools, which might have implications for data processing and analysis especially in the field of medical informatics.

The study by Murthy and Srilatha [4] is related to a comparative analysis of ML algorithms on a diabetes dataset. Although not focused on breast cancer, this reference underscores the broader application of machine learning in healthcare. It suggests that similar methodologies and algorithms employed in diabetes research could potentially be adapted and applied to breast cancer studies, highlighting the interdisciplinary nature of machine learning in medical research. The study [5] by Kaur and Kumari, published in Applied Computing and Informatics in July 2020, emphasizes on employing a machine learning based approach for predictive modelling and diabetic analysis. This paper likely explores the application of machine learning techniques to predict and analyse diabetes, emphasizing the significance of data-driven approaches in healthcare. The use of predictive modelling can contribute to early detection and intervention for better management of diabetes. The study by Katarya and Jain [6] discusses the comparison of various ML models for diabetes detection and is part of an IEEE conference publication from December 2020. The paper likely presents a comparative analysis of ML algorithms, evaluating their performance in detecting diabetes. Understanding the strengths and weaknesses of different models is essential for selecting the most suitable approach for accurate predictions.

In the study by Hassan et al. [7], the attention is directed toward predicting diabetes by employing an ensemble of various machine-learning classifiers. Published in a 2020 IEEE journal, the paper likely investigates the efficacy of ensemble methods, which involve combining multiple classifiers, to enhance the accuracy and resilience of models for predicting diabetes. Ensembling is a common strategy to enhance predictive performance by leveraging the diversity of individual classifiers. The study by Naz and Ahuja [8] introduces a deep learning based solution for diabetes prediction for the PIMA Indian dataset. Deep learning techniques, often associated with neural networks, can capture complex patterns in data and may offer advantages in predicting diabetes based on intricate relationships within the dataset. In the study by Patil and Ingle [9], featured in an IEEE conference publication dated June 2021, the paper engages in a comparative examination of diverse ML classification algorithms in the context of predicting diabetes, utilizing the Pima Indian Diabetics dataset.

In the study by written by Miao [10] the exploration revolves around employing ML algorithms for predicting diabetes, utilizing the PIMA Diabetes dataset. The paper likely delves into specific machine learning models and their performance in predicting the onset of diabetes, contributing to the ongoing research on diabetes prediction and highlighting the importance of accurate predictive analytics for early intervention. The research conducted by Abdulhadi and Al-Mousa [11], part of an IEEE conference publication from July 14, 2021, addresses diabetes detection using ML classification methods. This work likely extends the exploration of various classification techniques and their effectiveness in identifying diabetes cases. Such studies are crucial for understanding the strengths and weaknesses of different algorithms, aiding healthcare professionals in selecting appropriate tools for accurate diagnosis. In the study by Vaishali et al. [12], originating from an IEEE conference publication dated October 01, 2017, the paper introduces a feature selection method based on genetic algorithms and a MOE fuzzy classification algorithm, and both applied to the Pima Indians Diabetes dataset.

This work likely focuses on enhancing the predictive power of machine learning models by employing genetic algorithms for feature selection and leveraging fuzzy classification, displaying a multidimensional approach to diabetes prediction. The study by Lakhwani et al. [13], part of an IEEE conference publication from December 01, 2020, discusses the prediction of the onset of diabetes using ANN. Artificial neural networks, inspired by the human brain's structure, are powerful tools for capturing complex patterns, and this study likely explores their efficacy in predicting diabetes onset based on the available dataset.

Transitioning to cardiovascular health, references fifteen through twenty cover various aspects of heart disease

prediction using machine learning. The studies conducted by Mohan et al. [14], Ahmad et al. [15], and Bertsimas et al. [16] all explore different ML techniques for effective heart disease classification, highlighting the importance of accurate and timely diagnosis in cardiovascular health. The research by Li et al. [17] specifically addresses heart disease identification in the context of e-healthcare, emphasizing the integration of machine learning into healthcare systems for improved disease management. The two research studies by Pouriyeh et al. [18] and Kavitha et al. [19], both conference publications, collectively contribute to an extensive exploration and comparison of techniques within the realm of heart disease. These works highlight the variety of approaches and methodologies applied in this field.

In the realm of diabetic retinopathy analysis, the study by Roychowdhury et al. [20] lays the groundwork for leveraging machine learning to detect diabetic retinopathy. This early work likely explores the application of machine learning algorithms to analyse retinal images for signs of diabetic retinopathy, contributing to the development of automated diagnostic tools for diabetic patients. The study by Reddy et al. [21], an IEEE conference publication from February 2020, presents an ensemble-based ML model designed for classifying diabetic retinopathy. This paper likely extends the exploration of machine learning techniques in diabetic retinopathy analysis, emphasizing the use of ensemble methods that combine multiple models to enhance classification accuracy. Such approaches are critical in improving the reliability of diagnostic systems. Moving to a deep learning focus, the study by Qummar et al. [22], published in IEEE Journals & Magazine in 2019, explores a deep learning based approach for diabetic retinopathy detection. This work likely delves into the use of deep neural networks and ensemble methods to achieve robust and accurate detection of diabetic retinopathy, highlighting the potential of advanced techniques in the field. Expanding the focus to a comprehensive outlook on diabetes diagnosis, this research study by Solares et al. [27] carries out a comparative examination of various deep neural architectures concerning electronic health records. While not specifically focused on diabetic retinopathy, this review likely provides insights into the uses of deep learning in broader healthcare contexts, potentially offering comparative perspectives on the efficacy of different deep neural network architecture.

III. METHODOLOGY

In this extensive healthcare analytics project, we've compiled five diverse datasets spanning various health domains. These datasets include information on cancer, Pima diabetes, diabetic retina conditions, heart health, and diabetes in general. Each dataset has undergone thorough preparation for analysis, ensuring its usability and relevance in the respective healthcare domains. Data preprocessing has been applied to address missing values, employing appropriate methods like imputation or removal.

The methodology for building a machine-learning model typically involves the following steps:

Problem Definition: The initial phase involves clearly articulating the problem and specifying the desired outcome. This lays the foundation for understanding the type of machine learning model required and the necessary data for model training.

Data Collection and Preparation: Following problem definition, the subsequent step is the collection and preparation of data intended for model training. This encompasses data cleaning, addressing missing values, and formatting the data appropriately for modelling purposes.

Exploratory Data Analysis (EDA): During this stage, the gathered data undergoes analysis to discern its features, patterns, and relationships. This aids in identifying potential issues with the data and informs decisions about which features to include or exclude in the model.

Feature Engineering: Building on the insights gained from EDA, new features can be generated or existing ones transformed to more effectively represent the problem, thereby enhancing the model's performance.

Model Selection: With the prepared data, the next step involves selecting a suitable machine-learning model based on the type of the problem at hand. It takes into account the size and type of data, the desired accuracy, and computational resources. We have used the following supervised learning algorithms for the dataset with target variables.

- Linear Regression
- Logistic Regression
- Poisson Regression
- Decision Tree
- Support Vector Machine

In this comprehensive healthcare analytics endeavour, five distinct datasets related to various health domains have been gathered, including the Cancer dataset, Pima diabetic dataset, Diabetic retina dataset, heart dataset, and Diabetic dataset. Each dataset has been meticulously prepared for analysis, ensuring usability and relevance within their respective healthcare domains. Data pre-processing has been undertaken for each dataset, encompassing the identification and treatment of missing values through appropriate methods such as imputation or removal. Additionally, normalization techniques have been applied to ensure uniform scales among features within each dataset. The subsequent stage involves the segmentation of the data, wherein each dataset undergoes a division into two subsets: a training set comprising 70% of the data, and a testing set encompassing the remaining 30%. This separation facilitates the model's learning process on the training set and its evaluation on the testing set to assess its performance.

Feature selection and engineering strategies have been applied to each dataset, emphasizing domain-specific approaches that entail generating new features linked to healthcare measurements. Additionally, unsupervised feature selection methods identify the most relevant features within each dataset. The next stage includes selecting a model,

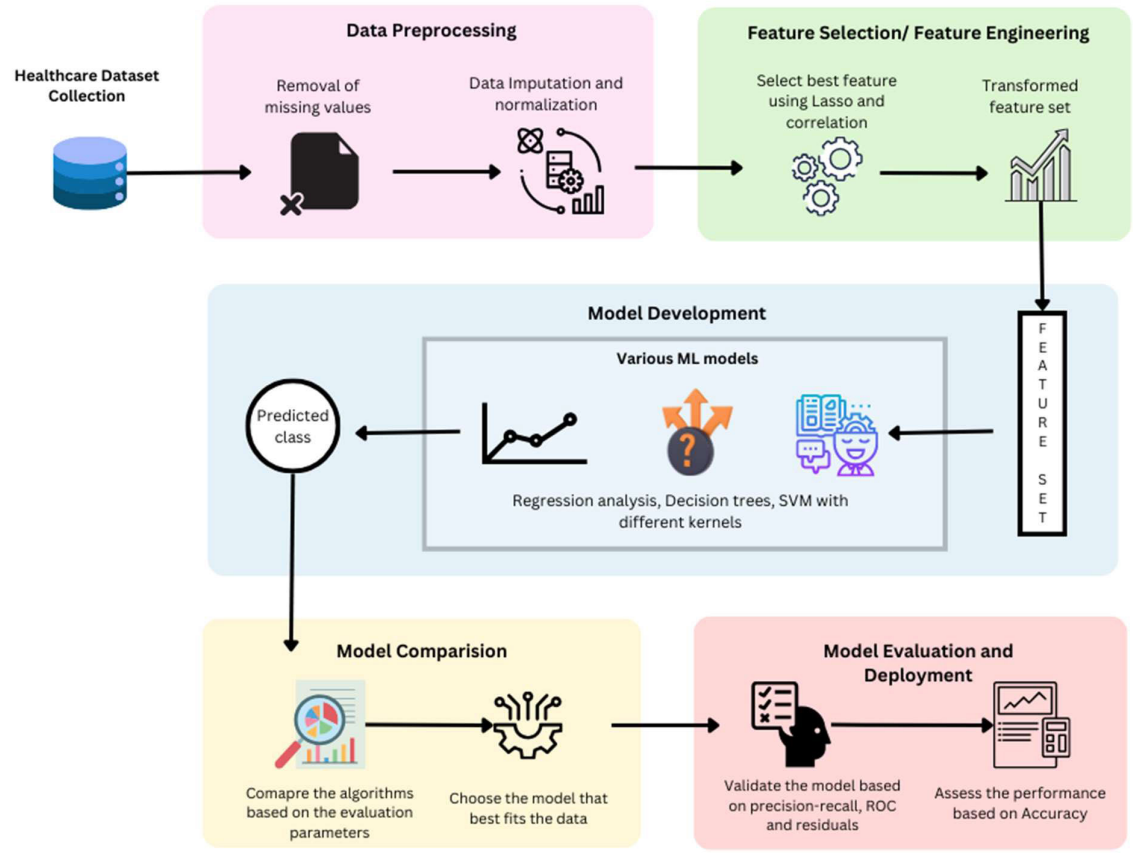

FIGURE 1. Architecture diagram.

where a variety of ML algorithms such as Decision Trees, Logistic Regression, SVM, Random Forest, and Gradient Boosting has been chosen due to their suitability for healthcare datasets. Hyperparameter tuning, carried out through grid search or random search techniques, guarantees optimal model performance for each dataset and the chosen algorithm.

In machine learning, hyperparameters are predefined parameters that are established before training a model, influencing how the model learns. For decision trees, two critical hyperparameters are maximum depth and minimum samples per leaf. Maximum depth dictates the depth of the tree, influencing its complexity and potential to over fit the training data. A deeper tree may capture intricate patterns in the data but risks memorizing noise, while a shallower tree may generalize better to unseen data. Minimum samples per leaf determines the minimum number of samples required to be present at a leaf node, regulating the granularity of splits. Higher values mitigate overfitting by ensuring each leaf represents a more significant portion of the data, enhancing generalization performance. However, excessively high values may lead to under fitting, resulting in oversimplified models.

On the other hand, Support Vector Machines (SVMs) boast hyperparameters such as regularization parameter (C) and kernel parameters (gamma for RBF kernel and degree for polynomial kernel). The regularization parameter, C, governs the trade-off between achieving a wider margin and minimizing classification errors. Larger C values prioritize correct classification of training instances, potentially leading to narrower decision boundaries and increased risk of overfitting. Conversely, smaller C values promote wider margins at the expense of higher training error tolerance, potentially enhancing generalization performance. Additionally, kernel parameters play a crucial role in non-linear SVMs by transforming input features into higher-dimensional space. Gamma, in the context of the RBF kernel, regulates the influence of individual training samples, with smaller values promoting smoother decision boundaries and higher values allowing more intricate decision boundaries. Degree, specific to polynomial kernels, determines the degree of the polynomial function used for kernel transformation, affecting the flexibility of decision boundaries. Higher degrees offer more complex decision boundaries, but excessive values may lead to overfitting, necessitating careful tuning.

In practice, the selection of hyperparameters for decision trees and SVMs hinges on the dataset characteristics and the trade-offs between model complexity and performance. For decision trees, striking a balance between depth and granularity is crucial, ensuring the model captures essential patterns while avoiding overfitting. Similarly, SVMs demand careful consideration of C and kernel parameters to achieve

optimal margins and decision boundaries without sacrificing generalization. Cross-validation techniques such as k-fold cross-validation facilitate the exploration of hyperparameter space, enabling the identification of the most suitable values for the given dataset. By understanding the roles and implications of these hyperparameters, researchers can fine-tune decision tree and SVM models effectively, maximizing their predictive power and applicability to real-world problems.

Following hyperparameter optimization, the models are trained on their respective training datasets, paving the way for comprehensive model evaluation on the corresponding testing datasets. Evaluation metrics specific to healthcare, such as accuracy, F1-score, precision, recall, and AUC-ROC, are employed to gauge model performance. Comparative analysis is then undertaken, discerning the strengths and weaknesses of each model on every healthcare dataset, ultimately pinpointing the best-performing model for each.

Accuracy evaluates the percentage of accurately classified instances relative to all instances within the dataset, offering an overall measure of model correctness. However, in healthcare applications with imbalanced datasets, precision, recall, and F1-score offer more nuanced assessments. A balanced assessment of a model's performance that takes into account both false positives and false negatives is provided by the F1-score, which is the harmonic mean of precision and recall. Recall gauges the model's capacity to identify every positive occurrence in the dataset, whereas precision assesses the accuracy of positive predictions. AUC-ROC also calculates the area under the curve that is between the true positive rate and the false positive rate, which sheds light on how well the model can discriminate. By analysing these metrics comprehensively, researchers gain a robust understanding of the models' performance and can select the best-performing model for each healthcare dataset, ensuring optimal predictive accuracy.

With the selection of the optimal model for each dataset, the subsequent steps involve model deployment and continuous monitoring in a production environment. Rigorous measures are implemented to seamlessly integrate the selected models into the healthcare system, ensuring their practical applicability. A monitoring system is established to track the ongoing performance of deployed models, with regular updates undertaken to maintain accuracy and relevance.

In parallel, ethical considerations are paramount throughout this process, with each dataset undergoing a thorough ethical review to ensure compliance with privacy and healthcare data guidelines. Measures are implemented to mitigate bias and guarantee fairness in predictions, upholding ethical standards in healthcare analytics. Finally, detailed documentation is crafted for each step of the analytical process for every healthcare dataset.

IV. DATASET DESCRIPTION

This section describes the datasets that have been used to build the models and the attributes involved. The following five datasets have been chosen and thoroughly analysed using different parameters and machine learning models. The results, inferences and conclusion of the same have been discussed in the later part of the report.

A. CANCER DATASET

A cancer dataset for machine learning is a collection of data from cancer cases in a population. It typically includes patient demographics, medical history, and specifics about the cancer diagnosis and treatment. The information can come from a variety of sources, including hospitals, clinics, and public health records. The goal of using this dataset in machine learning is to create models that can analyse the data and predict or recommend cancer diagnosis, prognosis, and treatment. A machine-learning model, for example, could be trained to identify risk factors for cancer, such as age, gender, family history, and lifestyle factors. An alternative model could be developed to forecast the probability of a patient developing breast cancer by analysing their medical history and demographic information. These models can be used to help healthcare providers make patient-care decisions and to identify populations at high risk for cancer. The machine learning cancer dataset can also be used to develop personalized treatment plans for patients based on the characteristics of their cancer and their personal health history. Researchers can gain insights into the causes of cancer and develop new prevention and treatment strategies by analysing patterns and relationships in data.

B. DIABETIC DATASET

The diabetic records and dataset for ML model includes data collected from individuals with diabetes. This dataset includes information about age, sex, fasting sugar, pp, sugar, cholesterol amongst other features, which are used to train the ML models. This predicts the likelihood of certain health outcomes, such as the development of diabetic complications. The goal is to use this information to improve the accuracy of diagnostic and treatment decisions, and ultimately improve the health outcomes of those with diabetes.

C. DIABETIC RETINA DATASET

The diabetic retina dataset covers information regarding the diagnosis and treatment of diabetic retinopathy, a disorder affecting the retina of the eye in diabetic patients. The dataset contains images of retinal regions of interest (ROI) that have been tagged with markers indicating damaged areas. In addition, the information contains measures of micro aneurysms (MA), which are minute, balloon-like enlargements in the blood vessels of the retina that can serve as an early predictor of diabetic retinopathy. This information holds significant importance in the early identification and treatment of diabetic retinopathy, utilized by healthcare professionals and researchers to enhance the accuracy of diagnostic and treatment decision-making. The diabetic retina dataset is an essential resource for machine learning models, as it provides a vast and diverse amount of data for training and evaluating algorithms.

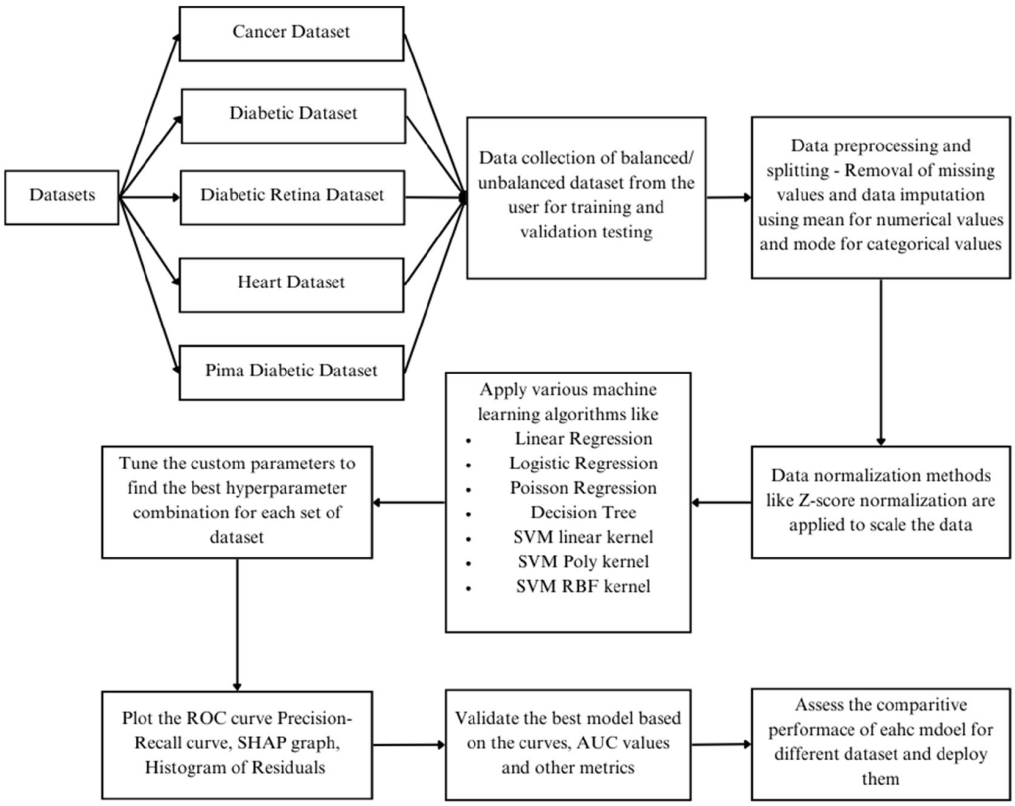

FIGURE 2. Methodology diagram to handle the datasets.

D. HEART DATASET

The heart dataset is a collection of medical information related to heart health. The dataset encompasses various variables instrumental in predicting the probability of heart disease, including thal (thalassemia), oldpeak, slope (peak exercise ST segment), exang (exercise-induced angina), and trest (resting blood pressure). These variables offer crucial insights into an individual's cardiac health, serving as inputs for training machine-learning models designed to forecast the likelihood of heart disease. Widely regarded as a valuable resource, the heart dataset proves essential for healthcare professionals and researchers, providing a comprehensive and diverse dataset for model training and evaluation. Enhancing the accuracy of these models entails considering additional factors like age, gender, and lifestyle elements, aiming to present a more holistic view of an individual's heart health.

E. PIMA DIABETIC DATASET

The Pima diabetic dataset comprises medical data associated with diabetes, featuring information on various variables for predicting the probability of diabetes. These variables include the number of pregnancies, skin fold thickness, plasma glucose concentration two hours after an oral glucose tolerance test and 2-hour serum insulin levels. These variables offer crucial insights into an individual's well-being and are utilized to train ML models for predicting the probability of diabetes. The Pima diabetic dataset serves as a valuable asset for healthcare professionals and researchers, furnishing an extensive and varied dataset for both model training and evaluation. Enhancing the precision of these models involves taking into account supplementary factors, including age, gender, and lifestyle elements, aiming to present a more comprehensive perspective on an individual's health.

V. STATISTICAL ANALYSIS AND INTERPRETATION

A critical stage in the construction of Machine Learning models is the pre-processing of datasets. This process entails cleansing and translating raw data into a format appropriate for modeling. Tasks such as managing missing values, handling outliers, modifying variables, and scaling data represent examples of pre-processing activities. By executing these tasks, the model can make predictions that are more accurate and generate outputs that are more precise. It is crucial to assess the performance of the chosen model using suitable metrics like accuracy, precision, and recall once the optimal model has been selected. Adjusting the model's hyper parameters then allows for its fine-tuning. The following table depicts and compares the accuracy achieved by using various ML models on the used datasets.

A. LINEAR REGRESION

Linear regression is a basic statistical technique widely employed in healthcare analytics for its ability to model and understand the relationships between variables,

making it an invaluable tool for healthcare professionals and researchers. In the healthcare domain, linear regression is used to analyse and predict various clinical and epidemiological outcomes, such as patient recovery times, disease progression, or the effectiveness of medical interventions. By fitting a linear model to data points, healthcare analysts can estimate the impact of specific factors on health outcomes, helping identify risk factors, determine treatment efficacy, and optimize resource allocation in healthcare settings. Furthermore, linear regression's simplicity and interpretability make it accessible to a wide range of healthcare stakeholders, enabling evidence-based decision-making and fostering data-driven advancements in patient care, health policy, and medical research. In an era of increasing data availability, linear regression remains a crucial analytical tool in the ever-evolving field of healthcare analytics. Equation (1) represents the assertions that in linear regression response is a linear function of the inputs.

$$y(x) = w^T x + \epsilon = \sum_{j=1}^{D} w_j x_j + \epsilon \tag{1}$$

In the given context, $w^T x$ denotes the inner or scalar product between the input vector x and the weight vector w^T of the model, while ϵ represents the residual error between our linear predictions and the actual response. Assuming the input is one-dimensional, we can express the anticipated response in the following manner.

$$\mu(x) = w_0 + w_1 x = w^T x \tag{2}$$

where w_0 represents the intercept or bias term, signifies the slope, and the vector x = (1, x) has been defined. Adding a constant term of 1 at the beginning of an input vector is a common notational technique, enabling the integration of the intercept term with the other model components. When w_1 is positive, it indicates an anticipation of the output to rise with an increase in the input.

B. LOGISTIC REGRESSION

Logistic regression is a powerful method extensively utilized in healthcare analytics to address classification and prediction tasks critical for patient care and medical research. Unlike linear regression, which is employed for continuous outcomes, logistic regression is specifically designed to handle binary or categorical outcomes, making it ideal for applications such as disease diagnosis, risk assessment, and treatment prediction. Healthcare professionals and researchers employ logistic regression to model the likelihood of an event occurring, such as a patient developing a specific condition or responding positively to a treatment. By analyzing a range of patient attributes, symptoms, and clinical variables, logistic regression allows for the identification of key predictors and their impact on medical outcomes. We can extend linear regression to the binary classification scenario by introducing two modifications. Initially, we substitute the Gaussian distribution for y with a more suitable Bernoulli distribution, especially when the response is binary the range of $y \in \{0, 1\}$. Secondly, we calculate a linear combination of the inputs, as in the linear regression case, but then we pass it through a function to ensure $0 \leq \mu(x) \leq 1$. This function is defined as,

$$\mu(x) = \text{sigm}(w^T x) \tag{3}$$

where, $\text{sigm}(\eta)$ denotes the sigmoid function, also recognized as the logistic or logit function. The sigmoid function is defined as,

$$\text{sigm}(\eta) = \frac{1}{1 + \exp(-\eta)} = \frac{e^{\eta}}{e^{\eta} + 1} \tag{4}$$

This information is invaluable for risk stratification, treatment selection, and the development of predictive models that enhance patient care, inform clinical decision-making, and contribute to advancements in healthcare, ultimately promoting more personalized and effective medical interventions. In the rapidly evolving landscape of healthcare analytics, logistic regression plays a crucial role in improving patient outcomes and driving evidence-based healthcare practices.

C. POISSON REGRESSION

Poisson regression is used to model count data, meaning data that takes on non-negative integer values, such as the number of patients diagnosed with a particular disease, or the number of doctor's visits per month. It is a type of generalized linear model (GLM), which means that it is a linear model where the dependent variable is transformed using a link function. The link function is used to ensure that the model's predictions are non-negative and fall within the range of possible values for the dependent variable. Poisson regression is a popular choice for modeling count data in healthcare analytics because it is relatively simple to implement and interpret. It can be used to answer a variety of questions, such as: What factors are associated with an increased risk of developing a particular disease? What is the average number of doctor's visits per month for patients with a particular condition? How effective is a new drug at reducing the number of hospitalizations for patients with a particular disease?.

D. DECISION TREE

Decision trees are characterized by the iterative division of the input space, creating distinct regions, and establishing a local model within each corresponding region of the input space. It is a powerful analytical tool in healthcare, employed to make informed decisions and extract valuable insights from complex medical data. These hierarchical, tree-like structures systematically split and categorize patient information, such as symptoms, medical history, and test results, to predict outcomes or aid in clinical decision-making. By iteratively partitioning the data based on the most discriminative attributes, decision trees can identify patterns, risk factors, or potential diagnoses. They offer transparency, making it easier for healthcare professionals to understand the reasoning behind each decision. In healthcare analytics, decision trees are commonly used to optimize treatment

plans, identify high-risk patient groups, and assist in disease diagnosis, ultimately improving patient care and resource allocation. Their versatility, interpretability, and ability to handle both categorical and continuous data make decision trees a valuable asset in navigating the complex landscape of healthcare data analysis.

Determining the optimal partitioning of data is a computationally challenging problem, as established by Hyafil and Rivest in 1976, rendering it NP-complete. Hence, a common approach is to employ a greedy procedure for computing a locally optimal Maximum Likelihood Estimate (MLE). This strategy is utilized by popular implementations such as CART, C4.5 and ID3. The split function in these implementations selects the optimal feature and corresponding value by following this process:

$$\begin{aligned}(j*, t*) = \text{argmin} j \in \{1, \ldots, D\} \\ \times \min t \in T_j \text{cost}\left(\{x_i y_i : x_{ij} \leq t\}\right) \\ + \text{cost}(\{x_i y_i : x_{ij} > t\})\end{aligned} \tag{5}$$

To construct a decision tree, we start with a data partition, D, comprising training tuples along with their respective class labels. The decision tree is built based on a set of candidate attributes, denoted as attribute_list. The pivotal element in this process is the attribute_selection method, which is responsible for determining the optimal splitting criterion. This criterion entails pinpointing the attribute that, when employed for division, leads to the most effective separation of data tuples into distinct classes. The process of splitting may include determining a splitting attribute and, in certain instances, either a split-point or a subset for subsequent partitioning. The generated decision tree serves as a hierarchical structure that reflects the relationships between attributes, facilitating the classification of new instances based on the learned patterns from the training data. The following steps depict the algorithm for the Decision Tree:

(1) Initialize a node N.
(2) If all tuples present in dataset D are of the same class C, designate node N as a leaf having class C and return.
(3) If the attr list is empty, then the leaf node is labelled as N with the majority class in D and return.
(4) Employ the Attribute Selection Method (ASM) on the datset D and attribute list to identify the optimal splitting criterion.
(5) Assign node N with the determined splitting criterion.
(6) In the case where the splitting attribute is discrete-valued and allows for multiway splits, modify attribute_list by excluding the splitting_attribute.
(7) For every outcome j of the chosen criterion:
(8) Define Dj as the data tuples in D corresponding to j.
(9) If Dj has no elements, append a leaf labeled with the majority class in D to node N.
(10) Otherwise, attach the node obtained by recursively applying the decision tree algorithm on Dj and the updated attribute_list.
(11) Return node

VI. VISUALIZATION ANALYSIS

A. CANCER DATASET

The Lasso feature selection model in Fig. 3 suggests the following features were more useful for prediction: Size, Shape, Nucl, Chro, and radius. It also includes the plots for linear regression, Poisson regression and Logistic regression. From the graph, it is clearly seen that Logistic and Poisson regression does not fit the curve.

The SHAP (SHapley Additive exPlanations) values are approximately equal across multiple attributes for two class variables (Class 0 and Class 1) indicates that, on average, each attribute is contributing similarly to the prediction for both classes.

By analyzing the correlation matrix and applying the Lasso feature selection model, we found out that the following features were more useful for prediction: Size, Shape, Nucl, Chro, radius. It also includes the plots for linear regression, Poisson regression and Logistic regression. From the graph it is clearly seen that Logistic and Poisson regression doesn't fit the curve. The k-means clustering analysis with k =3 and a distillation score below 2500 suggests that the data can be meaningfully partitioned into three distinct clusters, and the clusters exhibit good separation. This could be valuable information for understanding patterns or segments within the dataset. The SHAP (SHapley Additive exPlanations) values are approximately equal across multiple attributes for two class variables (Class 0 and Class 1) indicates that, on average, each attribute is contributing similarly to the prediction for both classes. This equality in SHAP values implies that the influence of the individual features on the model's output is balanced and not biased towards one particular class.

An ROC curve, accompanied by an AUC of 0.97, signifies that the models successfully distinguish between positive and negative instances. An AUC approaching 1 indicates a high true positive rate and a low false positive rate, indicating outstanding performance by the classifier.

From the Decision tree diagram, we can see that the maximum depth is limited to 5 and the minimum number of sample leaf nodes is 3. An accuracy of 0.97 implies that the models correctly classify 97% of instances, highlighting their overall effectiveness in predicting outcomes. Furthermore, precision-recall curves approaching a value of one indicate that the models achieve high precision and recall. High precision signifies a low rate of false positives, and high recall indicates capturing a large portion of actual positive instances. This balance is crucial, particularly in scenarios where false positives or false negatives have varying degrees of impact.

B. DIABETIC DATASET

Analysing Fig. 8, we can interpret from the Lasso feature selection model, that the following features were more useful for prediction: fastine, pp and sugar. It also includes the graphs for linear and passion regression and it can be noted that the linear regression graph does not fit the curve.

By analyzing the correlation matrix and applying the Lasso feature selection model, we found out that the following

TABLE 1. Table of comparison for accuracy score.

ML models	Datasets				
	Cancer Dataset	*Diabetic Dataset*	*Diabetic Retina Dataset*	*Heart Dataset*	*Pima Diabetic Dataset*
Logistic Regression	69.963	54.032	76.521	79.629	66.984
Linear Regression	95.970	91.935	68.260	82.407	73.333
Poisson Regression	20.695	46.774	65.760	68.518	64.761
Decision Tree	97.080	97.333	62.770	86.419	78.481
SVM - Linear kernel	97.080	97.333	70.129	80.246	79.746
SVM - Poly kernel (Deg)	97.080 (3)	97.333 (3)	61.038 (2)	82.716 (1)	77.215 (2)
SVM - RBF kernel	96.350	93.333	63.203	80.246	77.215

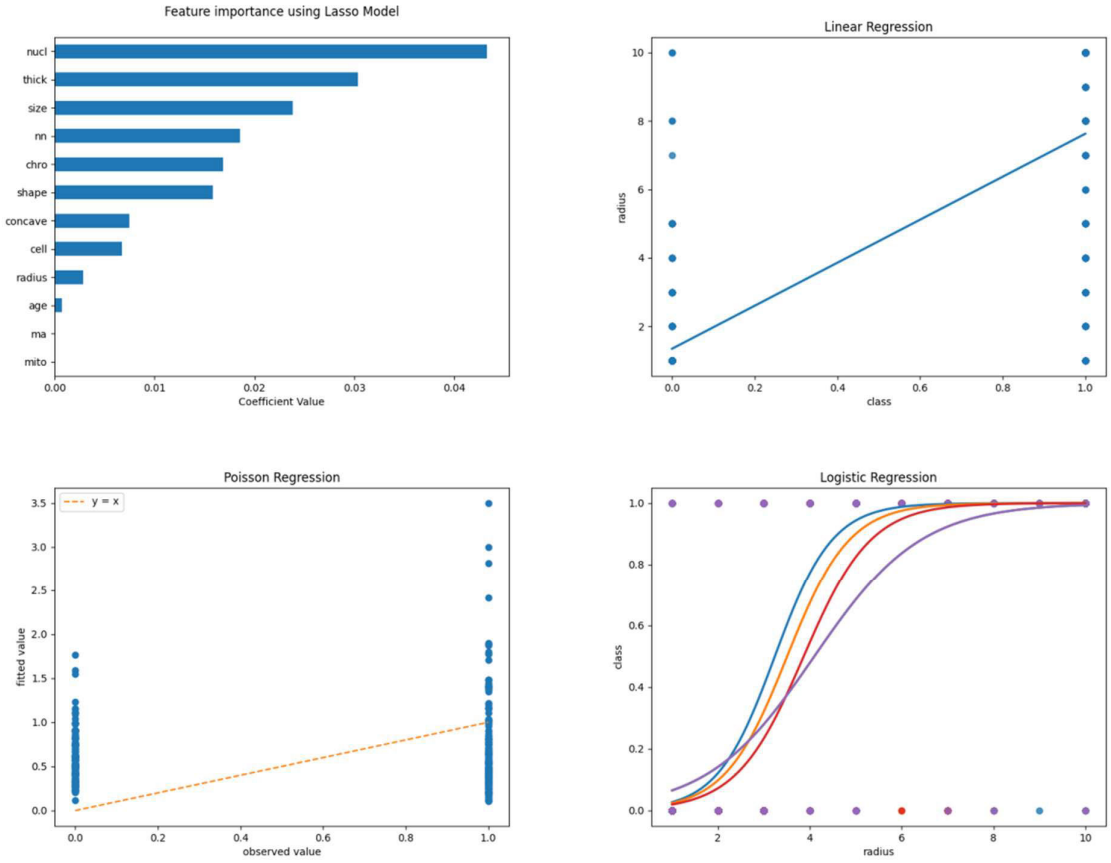

FIGURE 3. Feature selection and regression for cancer dataset.

features were more useful for prediction: Fastine, pp and sugar. It also includes the plots for linear regression and Poisson regression. From the graph, it is clearly seen that linear regression does not fit the curve. From the Decision tree, we can see that the maximum depth is one, the minimum number of sample leaf nodes is two, and the accuracy is higher. The exceptionally high ROC score of 0.99 and Precision-Recall score of 0.99 for the Decision Tree model

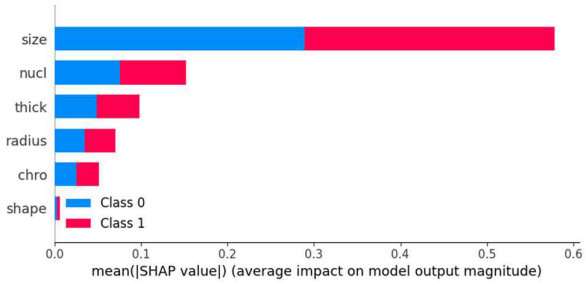

FIGURE 4. SHAP graph for the target variable to check for imbalance.

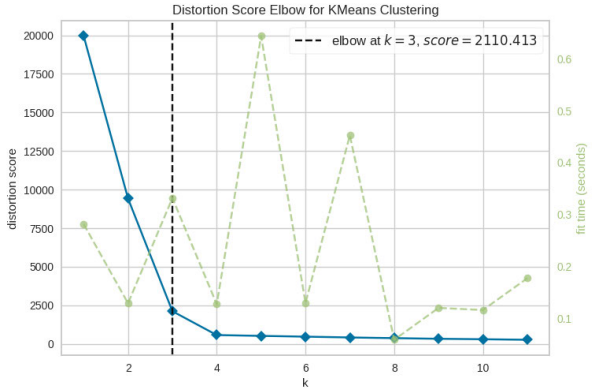

FIGURE 5. K-means clustering – Elbow graph for cancer dataset.

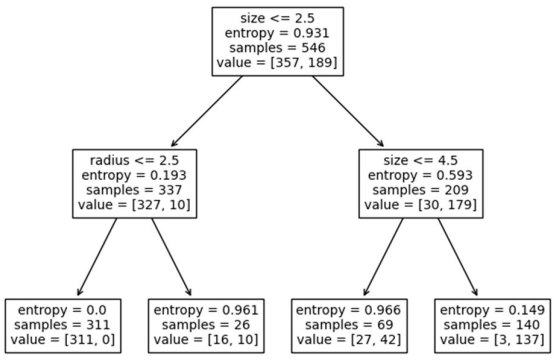

FIGURE 6. Decision tree for cancer dataset.

with hyperparameter tuning signify outstanding performance in discriminating between positive and negative instances. These statistics indicate that the model has achieved near-perfect classification, exhibiting an incredibly low false positive rate, high true positive rate, and precision. The model's overall accuracy of 97.33% further supports its proficiency in making accurate predictions across both classes.

C. DIABETIC RETINA DATASET

Upon examining the correlation matrix and employing the Lasso feature selection model in Fig. 10, it was determined that certain features, namely ma1, roi5, retinal, amfm, ma2, and ma4, exhibited greater utility for predictive purposes. Additionally, visual representations such as linear regression,

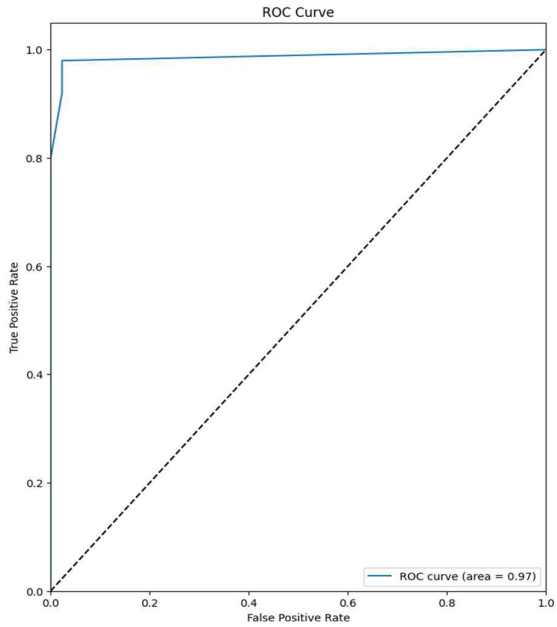

FIGURE 7. ROC curve for cancer dataset.

Poisson regression, and linear regression plots were generated. Notably, the graphical analysis revealed that the linear regression model did not align well with the observed curve.

By analyzing the correlation matrix and applying the Lasso feature selection model, we found out that the following features were more useful for prediction: ma1, roi5, retinal, amfm, ma2, ma4. It also includes the plots for linear regression, Poisson regression and linear regression. From the graph it's clearly seen that linear regression doesn't fit the curve.

From the Decision tree diagram, we can see that the depth of the tree is limited to four and the minimum number of sample leaf nodes is three. This improves the accuracy of the model. Given the critical nature of medical diagnoses, it is essential to carefully consider factors such as the sensitivity and specificity of the model.

A balanced Precision-Recall AUC implies that the model maintains a good equilibrium between correctly identifying positive cases and avoiding false positives. For the diabetic retinopathy database, the logistic regression model exhibits promising performance with a ROC of 0.83, a Precision-Recall value of 0.87, and an accuracy of 76.92%. These metrics suggest that the model is effective in distinguishing between instances of diabetic retinopathy and non-retinopathy cases. However, the assessment of whether it qualifies as a "good" or "bad" model depends on the specific goals and requirements of the application in the context of diabetic retinopathy.

The logistic regression plot includes the following features: Retinal, ma1, ma2, ma3, ma4, roi6, roi7, roi8, amfm. Form the regression curve for the Diabetic Retina Dataset, ma1, ma2, ma3, ma4, retinal and amfm looks to be more fitting the given dataset.

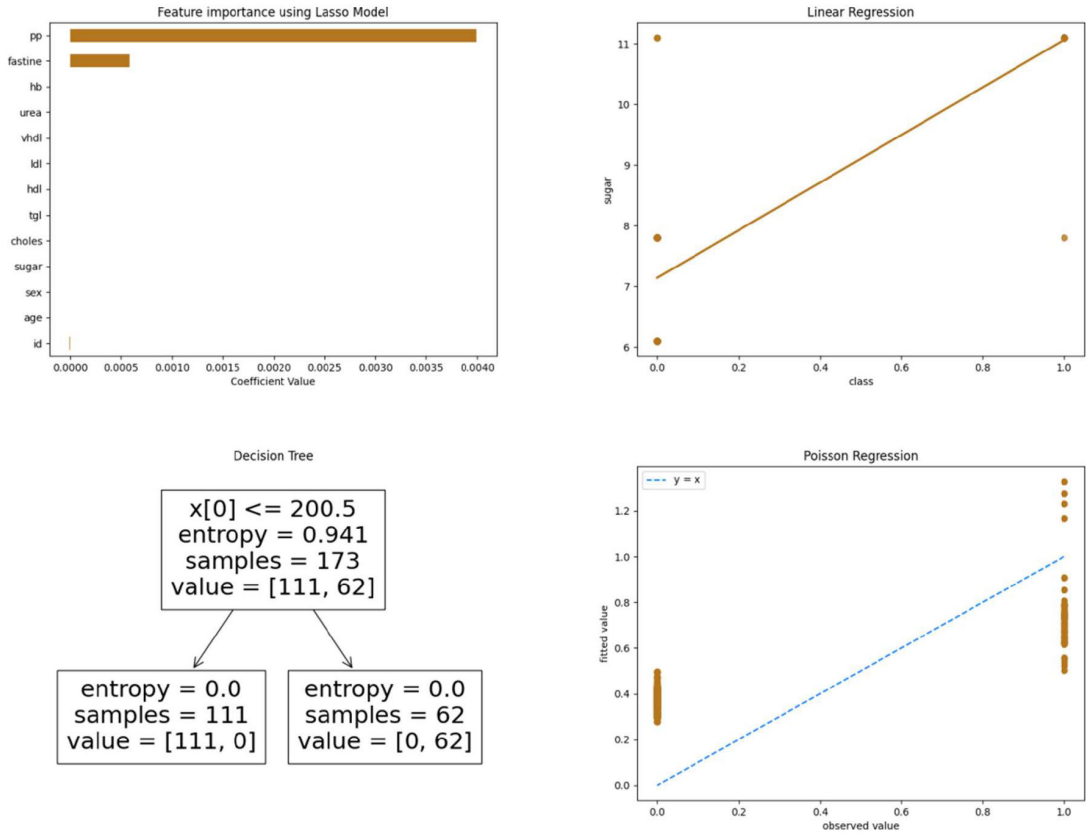

FIGURE 8. Feature selection, Regression and decision tree for diabetic dataset.

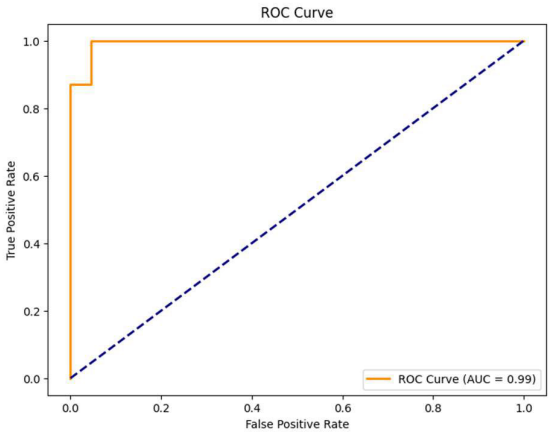

FIGURE 9. ROC curve for diabetic dataset.

D. HEART DATASET

Upon examining the correlation matrix and employing the Lasso feature selection model in Fig. 15, it was determined that certain features, namely exang, ca, cp, thal, bfs, exhibited greater utility for predictive purposes.

Additionally, visual representations such as linear regression, Poisson regression, and linear regression plots were generated. Notably, the graphical analysis revealed that the linear regression model and linear regression gave the same accuracy.

By analyzing the correlation matrix and applying the Lasso feature selection model, we found out that the following features were more useful for prediction: exang, ca, cp, thal, bfs. It also includes the plots for linear regression, Poisson regression and Logistic regression. From the graph, it is clearly seen that Linear and Logistic almost give the same accuracy.

From the Decision tree diagram, we can see that the maximum depth is limited to five and the minimum number of sample leaf nodes is three. The decision tree model for the heart dataset demonstrates strong performance with an ROC of 0.90 and precision-recall score of 0.78 suggesting good precision and recall trade-offs. However, it is important to note that while the ROC AUC is high, precision-recall AUC is slightly lower, indicating potential class imbalance or differing costs of false positives and false negatives. The precision-recall graph would provide insights into the trade-offs between precision and recall, with a higher precision desired for certain applications. As for the histogram of residuals, the concentration of residuals around 0 indicates that the model predictions are generally accurate, but the spread in the range from -0.25 to $+0.25$ suggests some variability in prediction errors, which might warrant further investigation into specific patterns or outliers in the data.

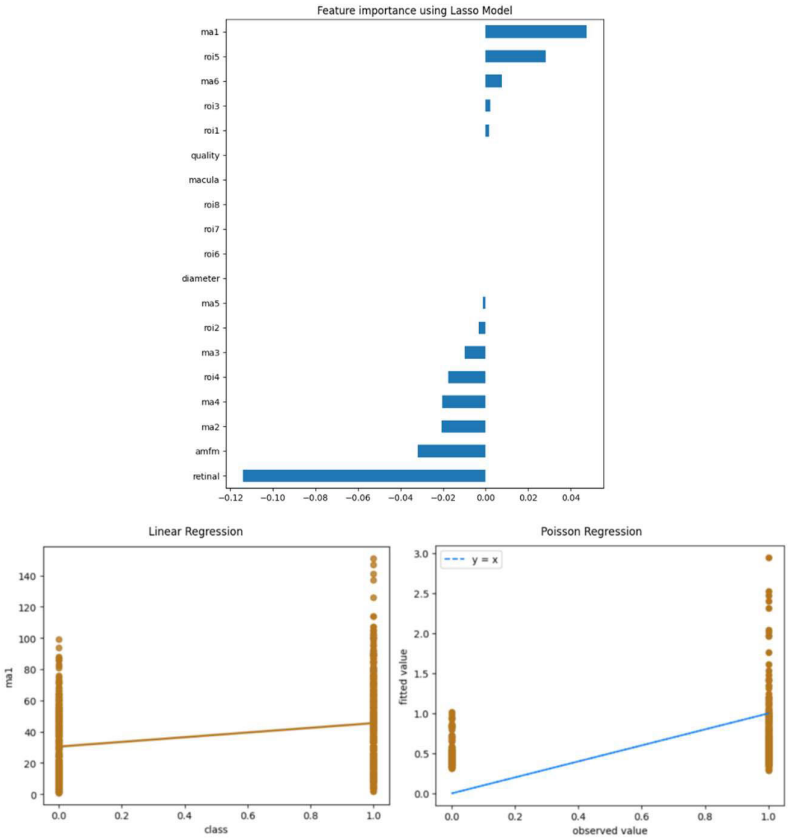

FIGURE 10. Feature selection using lasso and regression for diabetic retina dataset.

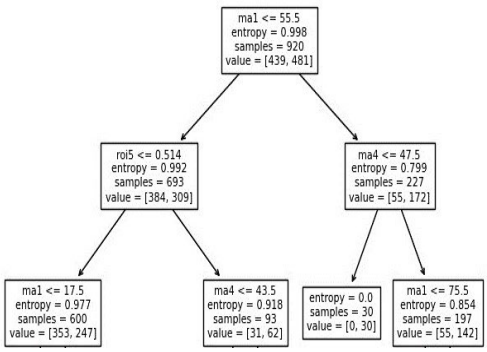

FIGURE 11. Decision tree for diabetic retina dataset.

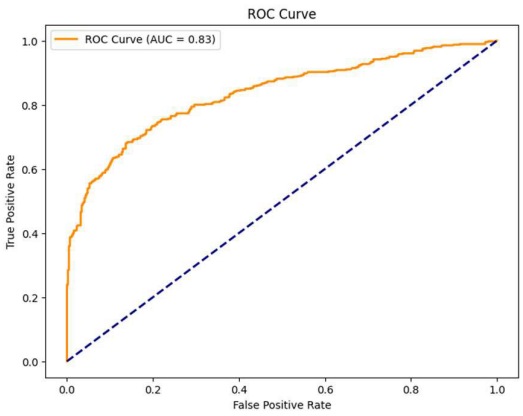

FIGURE 12. ROC curve for diabetic retina dataset.

E. PIMA DIABETIC DATASET

Through an analysis of the correlation matrix and the implementation of the Lasso feature selection model, it was identified that specific features—glucose, age, and insulin—proved to be more advantageous for predictive purposes. In addition, visual representations, including plots for linear regression, Poisson regression, and Logistic regression, were generated. Notably, the graphical analysis indicated that both Logistic and Poisson regression models did not align well with the observed curve.

From the Decision tree diagram, we can see that the tree's maximum depth is limited to five and the number of sample leaf nodes is three. This is done in order to improve the accuracy of the model. The performance metrics for the linear SVM model on the Pima Indian Diabetes dataset reveal a moderately effective classifier. The ROC of 0.82 and the precision-recall of 0.74, with a stepwise decrement in the curve, implies a trade-off between precision and recall. This

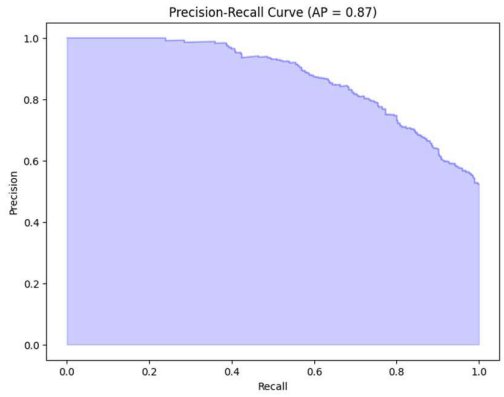

FIGURE 13. Precision-recall curve for diabetic retina dataset.

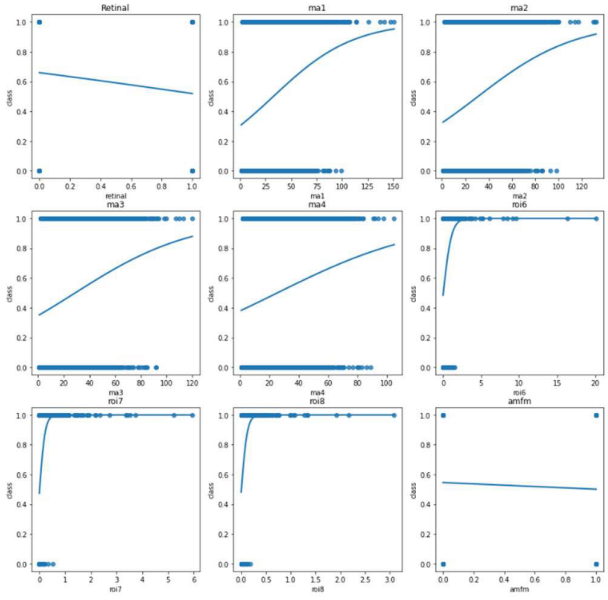

FIGURE 14. Logistic regression for diabetic retina dataset.

pattern may indicate challenges in achieving high precision without compromising recall, possibly due to imbalances in the dataset or complex decision boundaries. The model's overall accuracy of 79.746% suggests reliable generalization to the dataset, although the stepwise behavior in both ROC and precision-recall curves underscores the need for careful threshold selection to balance sensitivity and specificity in real-world applications.

VII. STATISTICAL ANALYSIS AND INTERPRETATION

A. CANCER DATASET

Linear regression is designed to predict continuous values, while logistic regression is designed to predict binary or multiclass outcomes. It's not accurate to say that linear regression is always more accurate than logistic regression for a classification problem. The accuracy depends on the specific problem and data, and it is possible that a logistic regression model could have better accuracy than a linear regression model for a classification problem, or vice versa. The accuracy of a linear regression model for a classification problem may be misleading, as the predictions from a linear regression model are continuous values, not discrete class labels. The continuous predictions must be threshold to generate binary or multiclass classifications. The choice of threshold can greatly affect the accuracy of the model. In addition, the SVM with linear and polynomial kernels yields the same result. If a dataset yields identical accuracy scores for both SVM model with a linear kernel and a SVM model with a polynomial kernel, several potential conclusions can be drawn:

- The data is linearly separable: This implies that the two classes are distinguishable by a straight line or hyperplane within the feature space. Consequently, a linear SVM with a linear kernel can attain equivalent accuracy to a polynomial SVM, as both models can identify a decision boundary effectively separating the two classes.
- The dataset exhibits a complex decision boundary: When the decision boundary is intricate, a polynomial SVM might be more apt for capturing the non-linear association between the features and the target. Nevertheless, it remains plausible that a linear SVM, with a meticulously selected set of features, can achieve comparable accuracy to a polynomial SVM.

TABLE 2. Table of comparison for ROC score.

Datasets	Metrics	
	ROC	*Precision-Recall*
Cancer dataset	0.97	96.27%
Diabetic dataset	0.99	97.33%
Diabetic Retina dataset	0.83	76.92%
Heart dataset	0.90	78%
Pima Diabetic dataset	0.82	79.7464%

B. DIABETIC DATASET

Here, we used K-Means clustering algorithm to predict the class variable since it is an unsupervised dataset. The dataset did not have any target variable so it has been generated using the K-means clustering algorithm. Linear regression is then designed to predict continuous values, while logistic regression is designed to predict binary or multiclass outcomes. It is not accurate to say that linear regression is always more accurate than logistic regression for a classification problem. It is possible that a logistic regression model could have better accuracy than a linear regression model for a classification problem, or vice versa. The accuracy of a linear regression model for a classification problem may be misleading, as the predictions from a linear regression model are continuous values, not discrete class labels. The continuous predictions must be thresholded to generate binary or multiclass classifications. The choice of threshold can greatly affect the accuracy of the model. If a SVM model with a linear kernel

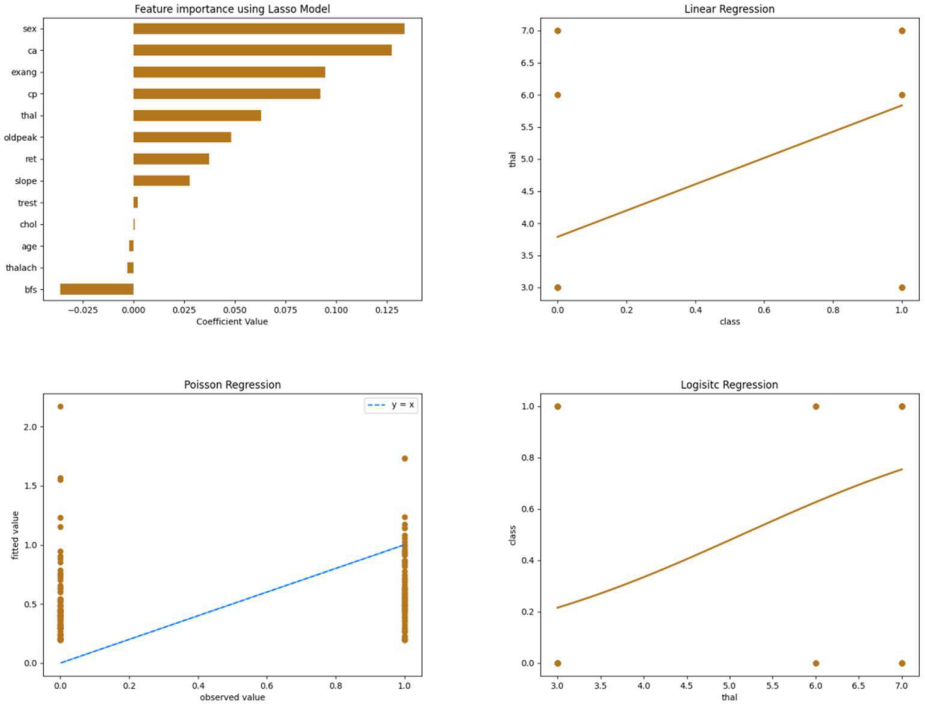

FIGURE 15. Feature selection and regression for heart dataset.

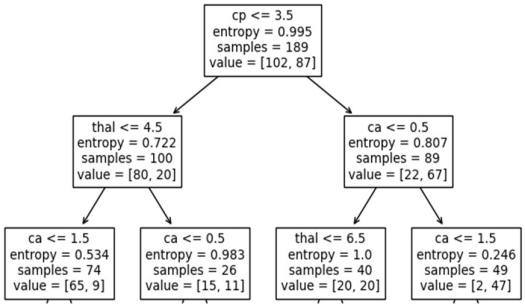

FIGURE 16. Decision tree for heart dataset.

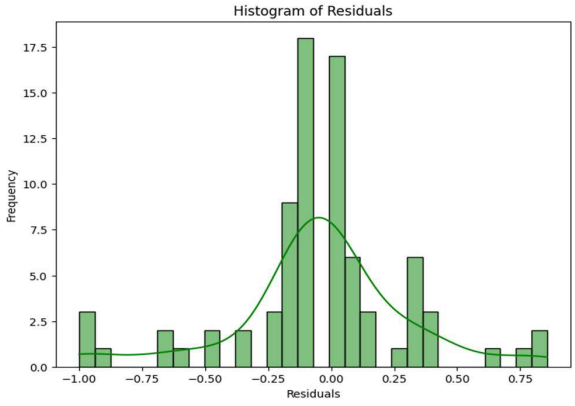

FIGURE 17. Histogram of residuals for heart dataset.

and another with a polynomial kernel both have the same accuracy score on a particular dataset then,

- The polynomial SVM is overfitting: It is possible that the polynomial SVM is overfitting to the training data and is not generalizing well to new, unseen data. In this case, a simpler linear SVM may be a better choice. It is always important to evaluate the performance of a model on a separate validation set to ensure that it is not overfitting.
- The dataset is small: If the dataset is small, then it is possible that the difference in accuracy between the linear and polynomial SVMs is not significant. In such cases, either model can be used without a significant impact on performance.

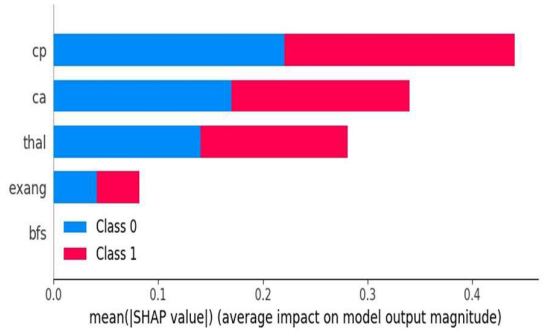

FIGURE 18. SHAP graph for the extracted features of heart dataset.

C. DIABETIC RETINA DATASET

The given problem is a multi-class classification problem and logistic, linear and Poisson regression models have been applied. On analyzing the results, it has been observed that,

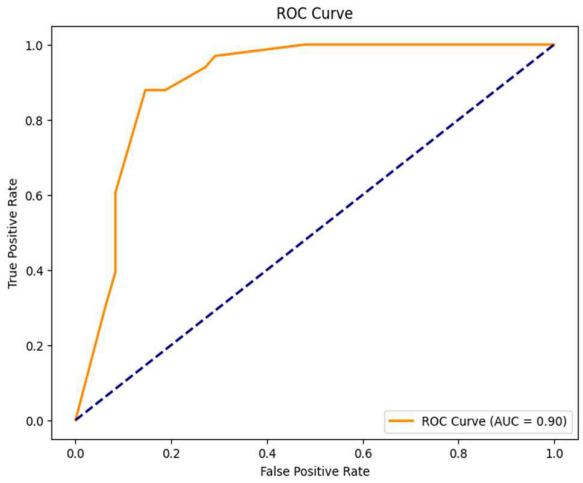

FIGURE 19. ROC curve for heart dataset.

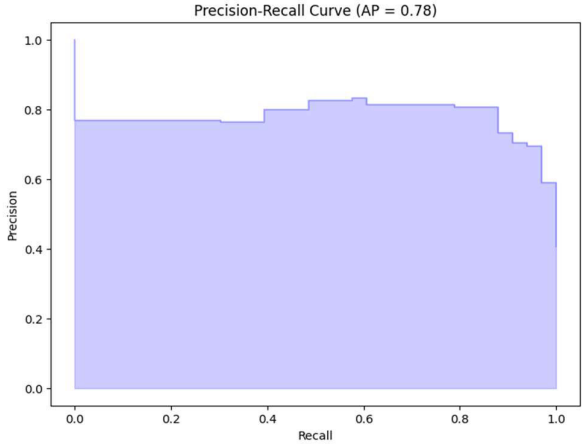

FIGURE 20. Precision recall curve for heart dataset.

logistic regression model proves to be the better model of the three. However, the accuracy scores of all the three models are not satisfactory. Reasons could be lower quality dataset, insufficient data to train the model. Since the features were scarcely correlated with the target variable, we used RFE based feature selection to choose the top features that could best fit the model. If a linear kernel SVM has a greater accuracy score than a SVM model with a polynomial kernel on a particular dataset, we can make a few possible inferences:

- The data is linearly separable and has a simple decision boundary: A linear SVM with a linear kernel can achieve a perfect separation, resulting in a higher accuracy than a polynomial SVM. Even if the data is not linearly separable and the decision boundary is simple, a linear SVM may be able to achieve a higher accuracy than a polynomial SVM.
- The feature space is not well-suited for a polynomial kernel: The polynomial kernel may not be a good fit for the particular feature space of the dataset. In such cases, using a linear kernel can be a better choice.
- The dataset is small: If the dataset is small, then the difference in accuracy between the linear and polynomial SVMs may not be significant. In such cases, either model can be used without a significant impact on performance.

D. HEART DATASET

The given problem is a multi-class classification problem and logistic, linear and Poisson regression models have been applied. On analyzing the results, it has been observed that, linear regression model proves to be the better model of the three. But, the accuracy scores of the other two models are not satisfactory. Reasons could be that the given data is not a Poisson distribution. It could be a normal distribution following a bell-shaped curve. Also, the Poisson regression is used for count variables mostly. The only feature that strongly correlated was 'thal' with the target variable. The level of thalassemia which is also a discrete variable with values 3,6,7. Therefore, we can say linear regression could also be used to train the model. If an SVM model with a linear kernel has a lower accuracy score than an SVM model with a polynomial kernel, it could mean that the dataset has a non-linear decision boundary. In such a case, the linear kernel may not be able to capture the complexity of the data and may under fit the data, leading to lower accuracy. The polynomial kernel, on the other hand, can model non-linear relationships between features. It does this by projecting the data into a higher-dimensional space where it can be separated by a linear boundary. This higher-dimensional space allows the polynomial kernel to capture more complex decision boundaries and could lead to a better accuracy score.

E. PIMA DIABETIC DATASET

The given problem is a multi-class classification problem and logistic, linear and Poisson regression models have been applied. On analyzing the results, it has been observed that, linear regression model proves to be the better model of the three. However, the accuracy scores of the other two models are not satisfactory. Reasons could be that the given data is not a Poisson distribution. It could be a normal distribution following a bell-shaped curve. In addition, the Poisson regression is used for count variables mostly. The only feature that strongly correlated was 'thal' with the target variable i.e. the level of thalassemia, which is also a discrete variable with values 3, 6, 7. In addition, we use only one target variable only so linear regression will give more accuracy than the other models. Therefore, we can say linear regression could also be used to train the model. If a Support Vector Machine (SVM) model with a linear kernel has a greater accuracy score than a SVM model with a polynomial kernel on a particular dataset, we can make a few possible inferences:

- The data has a simple decision boundary: If the dataset is linearly separable, a linear SVM with a linear kernel can achieve a perfect separation, resulting in a higher accuracy than a polynomial SVM.

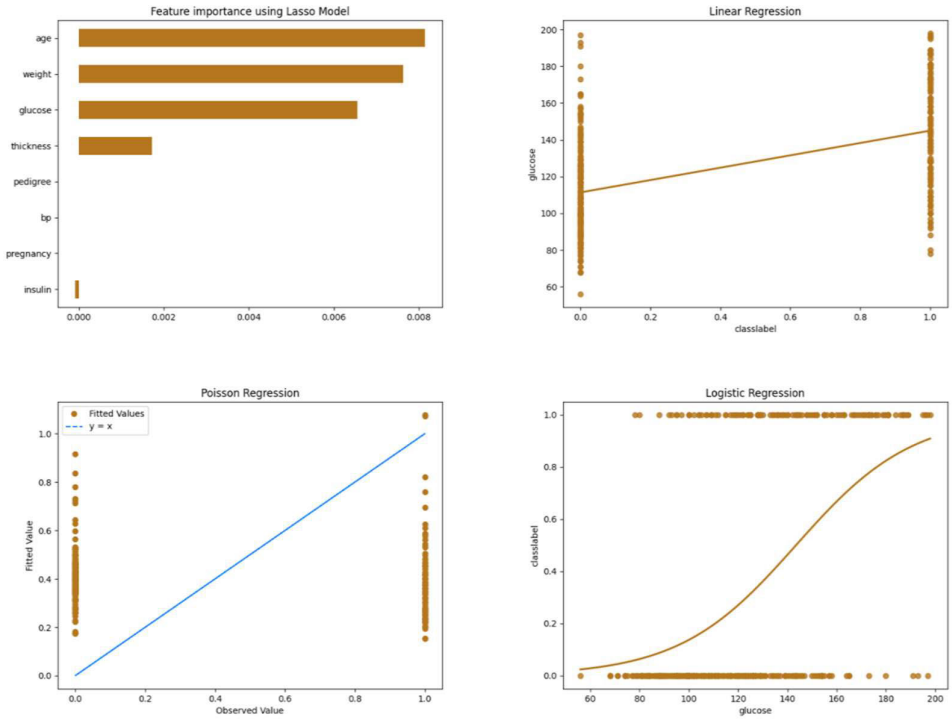

FIGURE 21. Feature selection and regression for pima diabetic dataset.

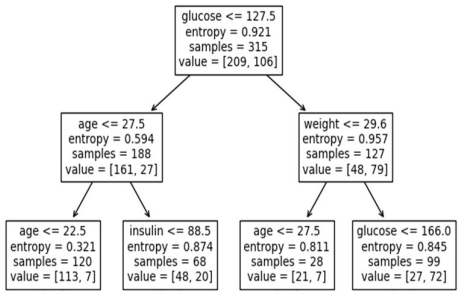

FIGURE 22. Dataset decision tree regression for pima diabetic dataset.

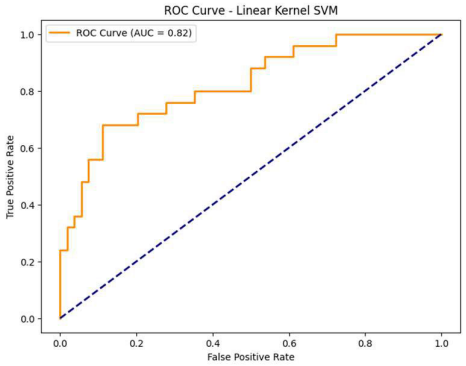

FIGURE 23. ROC curve for pima diabetic dataset.

- The feature space is not well-suited for a polynomial kernel: The polynomial kernel may not be a good fit for the particular feature space of the dataset. In such cases, using a linear kernel can be a better choice

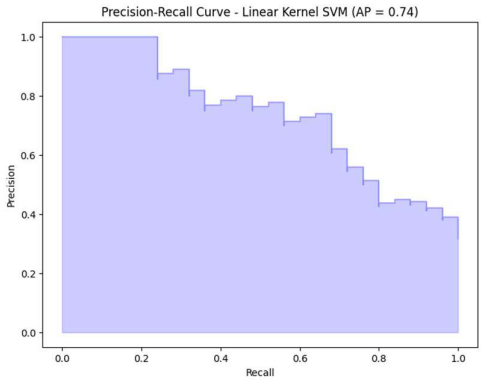

FIGURE 24. Precision-recall curve for pima diabetic dataset.

VIII. CONCLUSION

The research paper embarks on an extensive exploration of healthcare analytics, emphasizing its potential to transform patient care significantly. The machine learning models demonstrate impressive accuracy in predicting outcomes across various diseases, including cancer, diabetes, diabetic retinopathy, and heart-related conditions. Notably SVM and Decision Trees emerge as robust performers, achieving high accuracies in predicting cancer and diabetic outcomes, while Logistic regression shows notable success in predicting

diabetic retina outcomes, highlighting the need for diverse approaches in addressing different medical challenges.

Furthermore, the research delves into the complexities of feature selection and regression techniques, identifying critical factors that influence disease outcomes across datasets. Through meticulous analysis, significant predictors such as size, shape, and cholesterol levels are identified, providing valuable insights into the underlying mechanisms of disease progression. However, the comparative evaluation of machine learning models also reveals nuanced performance differences, emphasizing the importance of tailored algorithm selection based on dataset characteristics. Despite these challenges, the study underscores the transformative potential of healthcare analytics, facilitating early disease detection, personalized treatments, and improved patient outcomes.

To enhance the generalizability of the machine learning models, several measures have been implemented. Firstly, evaluation of the datasets was conducted to ascertain their representativeness of the broader patient population. This involved comprehensive analyses of demographics and clinical characteristics to validate the diversity of the data. Secondly, efforts were made to rectify any biases present in the dataset. Techniques such as oversampling and under sampling were utilized to address sampling biases or imbalances in class distribution. Through the evaluation of model performance metrics within each subgroup, any disparities were identified and necessary adjustments were made to promote fairness in our models' predictions. These measures collectively enhanced the reliability of our machine learning models, enhancing their potential for real-world applicability.

Looking ahead, the future of healthcare analytics holds immense promise for innovation and advancement. As machine-learning techniques evolve and integrate with emerging data sources such as genomics and wearable devices, opportunities for enhancing disease prediction and population health management will expand significantly. However, addressing ethical, regulatory, and technical challenges remains crucial to ensure the responsible and equitable use of healthcare analytics. This research lays the foundation for a future where precision medicine and personalized healthcare become standard practice, ultimately revolutionizing the healthcare landscape for the better. Additionally, it is important to note that accuracy alone is not sufficient for evaluating model performance, as metrics like precision, recall, and F1-score provide deeper insights. The choice of kernel in SVM models also depends on the data nature and problem context, necessitating careful evaluation and hyperparameter tuning to identify the most suitable model.

REFERENCES

[1] F. M. Agarap, "On breast cancer detection," in *Proc. 2nd Int. Conf. Mach. Learn. Soft Comput.*, Feb. 2018, pp. 1–18.

[2] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–11, Mar. 2019.

[3] S. Dalal, E. M. Onyema, P. Kumar, D. C. Maryann, A. O. Roselyn, and M. I. Obichili, "A hybrid machine learning model for timely prediction of breast cancer," *Int. J. Model., Simul., Sci. Comput.*, vol. 14, no. 4, pp. 1–24, Jun. 2022.

[4] G. D. Kimura, "A structure editor for abstract document objects," *IEEE Trans. Softw. Eng.*, volc. SE-12, no. 3, pp. 417–435, Mar. 1986.

[5] S. Murthy and J. Srilatha, "Comparative analysis on diabetes dataset using machine learning algorithms," in *Proc. 6th Int. Conf. Commun. Electron. Syst.*, 2021, pp. 1416–1422.

[6] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Appl. Comput. Informat.*, vol. 18, no. 1, pp. 90–100, Jul. 28, 2020, doi: 10.1016/J.ACI.2018.12.004.

[7] R. Katarya and S. Jain, "Comparison of different machine learning models for diabetes detection," in *Proc. IEEE Int. Conf. Adv. Develop. Electr. Electron. Eng. (ICADEE)*, Coimbatore, India, Dec. 2020, pp. 1–5, doi: 10.1109/ICADEE51157.2020.9368899.

[8] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[9] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, Apr. 14, 2020, doi: 10.1007/S40200-020-00520-5.

[10] V. Patil and D. R. Ingle, "Comparative analysis of different ML classification algorithms with diabetes prediction through pima Indian diabetics dataset," in *Proc. Int. Conf. Intell. Technol.*, Jun. 2021, pp. 1–9, doi: 10.1109/CONIT51480.2021.9498361.

[11] Y. Miao, "Using machine learning algorithms to predict diabetes mellitus based on PIMA Indians diabetes dataset," in *Proc. 5th Int. Conf. Virtual Augmented Reality Simulations*, Mar. 20, 2021, pp. 47–53.

[12] N. Abdulhadi and A. Al-Mousa, "Diabetes detection using machine learning classification methods," in *Proc. Int. Conf. Inf. Technol.*, Jul. 2021, pp. 350–354, doi: 10.1109/ICIT52682.2021.9491788.

[13] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE fuzzy classification algorithm on pima Indians diabetes dataset," in *Proc. Int. Conf. Comput. Netw. Informat. (ICCNI)*, Lagos, Nigeria, Oct. 2017, pp. 1–5, doi: 10.1109/ICCNI.2017.8123815.

[14] K. Lakhwani, S. Bhargava, K. K. Hiran, M. M. Bundele, and D. Somwanshi, "Prediction of the onset of diabetes using artificial neural network and pima Indians diabetes dataset," in *Proc. 5th IEEE Int. Conf. Recent Adv. Innov. Eng. (ICRAIE)*, Dec. 2020, pp. 1–6.

[15] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[16] G. N. Ahmad, H. Fatima, S. Ullah, and A. S. Saidi, "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151–80173, 2022, doi: 10.1109/ACCESS.2022.3165792.

[17] D. Bertsimas, L. Mingardi, and B. Stellato, "Machine learning for real-time heart disease prediction," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 9, pp. 3627–3637, Sep. 2021, doi: 10.1109/JBHI.2021.3066347.

[18] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in E-healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[19] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *Proc. IEEE Symp. Comput. Commun.*, Jun. 2017, pp. 204–207.

[20] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Coimbatore, India, Jan. 2021, pp. 1329–1333, doi: 10.1109/ICICT50816.2021.9358597.

[21] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "DREAM: Diabetic retinopathy analysis using machine learning," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 5, pp. 1717–1728, Sep. 2014, doi: 10.1109/JBHI.2013.2294635.

[22] G. T. Reddy, S. Bhattacharya, S. Siva Ramakrishnan, C. L. Chowdhary, S. Hakak, R. Kaluri, and M. Praveen Kumar Reddy, "An ensemble based machine learning model for diabetic retinopathy classification," in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng.*, Vellore, India, Feb. 2020, pp. 1–6, doi: 10.1109/IC-ETITE47903.2020.235.

[23] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. Ahmed Khan, and W. Jadoon, "A deep learning ensemble approach for diabetic retinopathy detection," *IEEE Access*, vol. 7, pp. 150530–150539, 2019, doi: 10.1109/ACCESS.2019.2947484.
[24] M. T. Islam, H. R. H. Al-Absi, E. A. Ruagh, and T. Alam, "DiaNet: A deep learning based architecture to diagnose diabetes using retinal images only," *IEEE Access*, vol. 9, pp. 15686–15695, 2021, doi: 10.1109/ACCESS.2021.3052477.
[25] Z. Liu, C. Wang, X. Cai, H. Jiang, and J. Wang, "Discrimination of diabetic retinopathy from optical coherence tomography angiography images using machine learning methods," *IEEE Access*, vol. 9, pp. 51689–51694, 2021, doi: 10.1109/ACCESS.2021.3056430.
[26] S. H. Abbood, H. N. A. Hamed, M. S. M. Rahim, A. Rehman, T. Saba, and S. A. Bahaj, "Hybrid retinal image enhancement algorithm for diabetic retinopathy diagnostic using deep learning model," *IEEE Access*, vol. 10, pp. 73079–73086, 2022, doi: 10.1109/ACCESS.2022.3189374.
[27] V. Raman, P. Then, and P. Sumari, "Proposed retinal abnormality detection and classification approach: Computer aided detection for diabetic retinopathy by machine learning approaches," in *Proc. 8th IEEE Int. Conf. Commun. Softw. Netw. (ICCSN)*, Jun. 2016, pp. 636–641, doi: 10.1109/ICCSN.2016.7586601.
[28] J. R. A. Solares, F. E. D. Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A. C. Pinho Gomes, A. H. Payberah, M. Zottoli, M. Nazarzadeh, N. Conrad, K. Rahimi, and G. Salimi-Khorshidi, "Deep learning for electronic health records: A comparative review of multiple deep neural architectures," *J. Biomed. Informat.*, vol. 101, Jan. 01, 2020, Art. no. 103337, doi: 10.1016/J.JBI.2019.103337.
[29] C.-Y. Fan, P.-C. Chang, J.-J. Lin, and J. C. Hsieh, "A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification," *Appl. Soft Comput.*, vol. 11, no. 1, pp. 632–644, Jan. 01, 2011, doi: 10.1016/J.ASOC.2009.12.023.
[30] L. Verma, S. Srivastava, and P. C. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data," *J. Med. Syst.*, vol. 40, no. 7, pp. 1–45, Jun. 11, 2016, doi: 10.1007/S10916-016-0536-Z.
[31] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.
[32] G. S. Bhavekar and A. D. Goswami, "A hybrid model for heart disease prediction using recurrent neural network and long short term memory," *Int. J. Inf. Technol.*, vol. 14, no. 4, pp. 1781–1789, Feb. 21, 2022, doi: 10.1007/S41870-022-00896-Y.
[33] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Syst. Appl.*, vol. 35, nos. 1–2, pp. 82–89, Jul. 01, 2008, doi: 10.1016/J.ESWA.2007.06.004.
[34] S. Kilicarslan, K. Adem, and M. Celik, "Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network," *Med. Hypotheses*, vol. 137, Apr. 01, 2020, Art. no. 109577, doi: 10.1016/J.MEHY.2020.109577.
[35] R. D. Howsalya Devi, A. Bai, and N. Nagarajan, "A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms," *Obesity Med.*, vol. 17, Mar. 01, 2020, Art. no. 100152, doi: 10.1016/J.OBMED.2019.100152.
[36] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," in *Proc. 8th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2017, pp. 386–390, doi: 10.1109/ICSESS.2017.8342938.
[37] B. Venkaiahppalaswamy, P. Prasad Reddy, and S. Batha, "Hybrid deep learning approaches for the detection of diabetic retinopathy using optimized wavelet based model," *Biomed. Signal Process. Control*, vol. 79, Jan. 01, 2023, Art. no. 104146, doi: 10.1016/J.BSPC.2022.104146.
[38] P. Ganesh and P. Sripriya, "A comparative review of prediction methods for pima Indians diabetes dataset," *Advances Intelligent Systems Computing*, 2020, pp. 735–750, doi: 10.1007/978-3-030-37218-7_83.
[39] M. S. Satu, S. T. Atik, and M. A. Moni, "A novel hybrid machine learning model to predict diabetes mellitus," in *Proc. Int. Joint Conf. Comput. Intell.*, Jan. 01, 2020, pp. 453–465.
[40] A. T. Alhasani, H. Alkattan, A. A. Subhi, E. S. M. El-Kenawy, and M. M. Eid, "A comparative analysis of methods for detecting and diagnosing breast cancer based on data mining," *J. Artif. Intell. Metaheuristics*, vol. 4, no. 2, pp. 08–17, 2023, doi: 10.54216/JAIM.040201.

A. SHEIK ABDULLAH (Member, IEEE) received the Ph.D. degree from Anna University. He is currently an Assistant Professor Senior with the School of Computer Science Engineering, Vellore Institute of Technology, Chennai. During his research, he was invited as a Visiting Faculty Member at the Institute of Mathematical Sciences (IMSC), Chennai, and contributed and involved in computational biology, mathematical decision support models in clinical informatics, data analytics, and statistics. He is also a Visiting Researcher with Chennai Mathematical Institute (CMI) and is engaged in works corresponding to timed automata and its applications. More recently, he has contributed his novelty in assessing risk factors that contribute to type II diabetes with swarm intelligence and machine learning approaches. His works correspond to real-time analysis of medical data and medical experts with the development of clinical decision support models for hospitals in rural areas. He also contributed his research intelligence in NLP, big data, knowledge-based systems, E-governance, learning analytics, and probabilistic planning algorithms. He has published more than 45 archival research articles to his credit, 15 book chapters, and a book. Since July 2015, he has been an Active Member of ACM and IEEE and being recognized for his extraordinal contribution to ACM chapter events with the ACM Faculty Sponsor Recognition (2015–2022). Being a Gold Medallist (PG), he has been awarded the Honourable Chief Minister Award for the best project in E-governance. He contributes his interests in various international conferences and serves as a Reviewer for IEEE, IEEE Access, Elsevier, Springer, and CRC publishers.

V. NAGA PRANAVA SHASHANK is currently pursuing the bachelor's degree in computer science and engineering with Vellore Institute of Technology, Chennai. Throughout his academic journey, he has consistently shown a deep interest in machine learning and deep learning. He has actively participated in various research projects and coauthored several research articles, with three already published. In one significant contribution, he innovatively analyzed technical indicators for stock market prediction, offering fresh insights in financial technology. Additionally, he has conducted a comprehensive analysis comparing the performance of machine learning and deep learning techniques in recommendation systems, providing valuable perspectives on their respective efficacies. Alongside his research endeavors, he engages actively in conferences, sharing his findings and collaborating with peers and experts. His innovative approaches and contributions exemplify the impactful role students can play in advancing computational sciences and technology.

D. ALTRIN LLOYD HUDSON is currently pursuing the bachelor's degree in computer science and engineering with Vellore Institute of Technology, Chennai, with a deep passion for machine learning, deep learning, and artificial intelligence. Throughout the academic journey, he has consistently demonstrated a keen interest in technology and a commitment to practical applications. Actively participating in a variety of projects, he has proven adept at translating theoretical concepts into tangible solutions. His academic achievements include the publication of two significant articles, which explore the novel use of machine learning in stock market prediction. Beyond the academic realm, he has contributed to the broader scientific community by participating in conferences and fostering valuable connections. An example of interdisciplinary work is the development of a music recommendation system for Spotify, highlighting his commitment to enhancing user experiences through advanced technological solutions. Positioned for a promising future, he is set to make meaningful contributions to the ever-evolving field of computer science and engineering.