Kannan *et al.*
Health Information Science and Systems (2026) 14:43
https://doi.org/10.1007/s13755-026-00438-x

Health Information Science and Systems

RESEARCH

An enhanced heart disease prediction model based on linear Diophantine fuzzy-integrated supervised machine learning

Jeevitha Kannan[1,2], Vimala Jayakumar[3], Nasreen Kausar[4,6] and Liang Kong[5*]

Abstract

The medical diagnosis often dealt with uncertainty and vagueness that hindered the effectiveness of conventional ML approaches. This limitation was overcome by the integration of the LDFS with ML algorithms in this study on heart disease diagnosis. The LDF framework is a powerful structure that has reference parameters that can easily change the physical meaning of its attributes. Our proposed hybrid model eliminates the need for pipelines like in conventional ML for handling categorical and numerical features, as it accommodates both feature types through membership functions. Several ML algorithms, like logistic regression, decision tree, support vector machine, and XGBoost, were evaluated on both the crisp dataset and LDF-based datasets. A comparative analysis demonstrates that our proposed LDF-ML consistently outperforms conventional ML algorithms in classifications. All the performance metrics were increased on LDF-Datasets by 0.97% in accuracy, 0.95% in precision, 0.99% in recall, and 0.97% in F1 score for the XGBoost Algorithm. Thus, the proposed integration provides a new direction for medical diagnosis as well as decision-making in terms of handling ambiguity with improved interpretability.

Introduction

Worldwide, cardiac disease is one of the most crucial global diseases which consistently increasing in the mortality rate. As per the reports of the WHO, heart disease causes 17.9 million deaths yearly, which is nearly 32% of all global deaths. Although there is a huge advancement in the healthcare structure and early diagnosis approaches [1]. Heart disease continues to pose a serious issue in the healthcare sector. Thus, improving the early and accurate diagnosis is the only solution to solve this issue. Prior detection of heart disorders plays a crucial role in reducing the fatality rates.

The traditional diagnosis system [2] mostly depends on clinical tests, scanning, and a physician's expertise. But these methods are often affected by incomplete data and inconsistencies that make the diagnosis uncertain. To overcome these challenges, ML has emerged as a powerful tool capable of identifying hidden patterns and correlations from large datasets. ML is one of the branches of Artificial Intelligence that primarily focuses on data analysis and improves performance in work without explicit programming. The potential of ML was enhanced by the combination of computational statistics and statistical learning through mathematical optimization tools and practical applications. ML algorithms analyze the data to identify patterns and make decisions to handle several complex tasks like classification, regression, and clustering effectively. There are several algorithms was demonstrated for prediction and classification purposes, such as LR, Random Forest, DT, SVM, K-Nearest Neighbour, XGB, etc. However, these conventional models operate on crisp numerical inputs, and their accuracy significantly decreases when faced with uncertain and incomplete information.

On the other hand, medical datasets often come with linguistic or qualitative attributes, namely type of chest pain, etc., that cannot be precisely quantified. In such scenarios, conventional ML fails to address the inherent ambiguity, which leads to misclassification. To solve

*Correspondence: lkong9@uis.edu
[5] Department of Mathematical Sciences and Philosophy, University of Illinois Springfield, Springfield, IL, USA
Full list of author information is available at the end of the article

this limitation, fuzzy-based models are introduced to enhance the handling of uncertainty in medical diagnosis with ML approaches [3]. FS can help to manage linguistic and uncertain concepts and also reduce confusion in healthcare decision-making procedures [4]. FS frameworks are mainly used in medical diagnosis to address complexity. The ML component uses large medical datasets to recognize complex patterns, and the FS interprets linguistic elements of medical knowledge [5]. By managing ambiguous or imprecise data, FS helps to improve the diagnostic process.

Literature review

FS was initially introduced by Zadeh to capture the uncertain environments [6]. A fuzzy system allows the feature to take partial membership between 0 and 1 rather than a binary classification. The membership function represents how much the patient belongs to a certain risk. Thus, the integration of fuzzy systems in ML is more important to capture the uncertain data more precisely [7]. This integration will get both the benefits of fuzzy and ML. Fuzzy Logic technologies are effective in addressing the fuzziness of medical diagnosis, as evidenced by their widespread use in medicine [8]. The integration of FS and ML approaches allows for the synergistic use of their respective strengths [9]. ML techniques improve the efficiency and speed of the decision process, while the FS framework is utilized for representing uncertain information [10]. This research highlights the importance of merging FS and ML to develop sophisticated approaches that help medical practitioners and improve patient care results.

Initially, FS and fuzzy logic were integrated with ML and applied in a wide range of applications. Later, as an extension of FS, the IFS framework was incorporated into ML [11], making the results clearer with an extra non-membership grade for each feature [12]. Also, the neutrosophic fuzzy-based dataset combined with an ML approach in lung cancer detection [13].

Uncertainty is represented by two components in intuitionistic and Pythagorean fuzzy sets, whereas three components are used in neutrosophic sets. But, LDFS offers more mathematical flexibility when compared to traditional fuzzy sets like FS, IFS, PFS [14, 15] due to the addition of reference parameters. These parameters provide independent control over membership and non-membership degrees, allowing for more precise uncertainty partitioning [16]. The term 'Linear Diophantine' represents the linear constraint that reflects the interaction between membership and non-membership degrees through reference parameters. Rather than requiring all information to fall within a strict unit interval, this structure intuitively permits uncertainty to be modeled within an enlarged yet regulated solution space. This leads to increased feature resolution, deeper representation of alternatives, and greater adaptability in complicated decision-making situations. Clustering analysis based on the Linear Diophantine FS has been introduced and applied to assess logistic efficiency [17]. The LDFS framework also finds broad applications, particularly in prioritizing risk factors [18]. In medical diagnosis, where symptoms frequently overlap, data may simultaneously support and contradict a condition [19, 20]. In such cases, the flexibility of LDFS is more crucial. While retaining mathematical consistency and interpretability, the Linear Diophantine formulation offers a rational method to account for this increased uncertainty.

This work proposes a model that can handle verbal and numerical inputs, making it accessible to non-specialists in healthcare. The main goals are to improve diagnostics, patient care, and resource use. Advancements in AI and ML, including disease prevention, personalized medicine, and digital diagnostics, highlight the need for high-quality decision support systems to address healthcare concerns and improve patient outcomes. This research study examines the use of LDF systems in healthcare, highlighting their ability to handle and express ambiguous information. Using fuzzy values can help medical practitioners manage linguistic nuances and make clearer decisions.

Motivation

The motivation behind this study is the possibility of better patient care outcomes through the fuzzy-based ML algorithms. These algorithms enable the detection of the disease in early stages. The hybrid idea can take partial truths and uncertainties. To overcome difficulties in medical diagnosis tasks, this study aims to create a sophisticated model that incorporates ML approaches, particularly in combination with LDF sets. Improving the accuracy in diagnosis and increasing the cure rate of patients and thereby optimizing resource use in medical diagnosis, are the primary objectives of this study. Thus, this study aims to integrate LDFSs in ML approaches for heart disease diagnosis.

Within this hybrid perspective, the Linear Diophantine Fuzzification (LDF) framework extends the expressive capacity of fuzzy representation through its reference parameters, allowing finer adjustment of uncertainty levels. In our preliminary exploration, the principal-component analysis (PCA) of the crisp dataset (Fig. 1) exhibited overlapping class distributions, indicating that the original features provided limited separability. In contrast, when the same data were transformed using LDF-based fuzzification, the resulting LDF-PCA projection (Fig. 2)

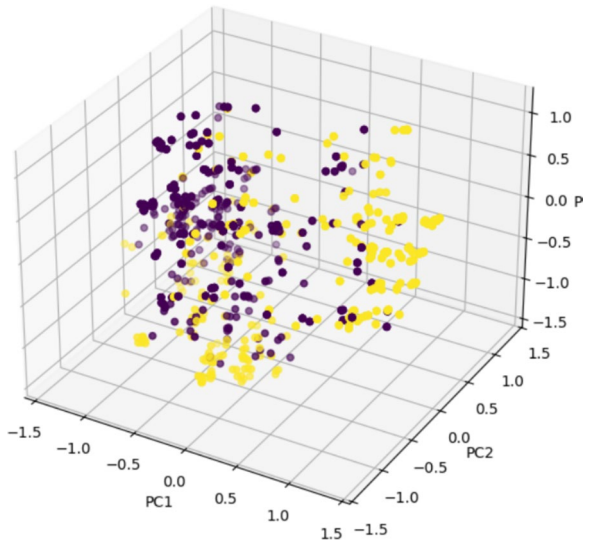

Fig. 1 PCA of crisp dataset

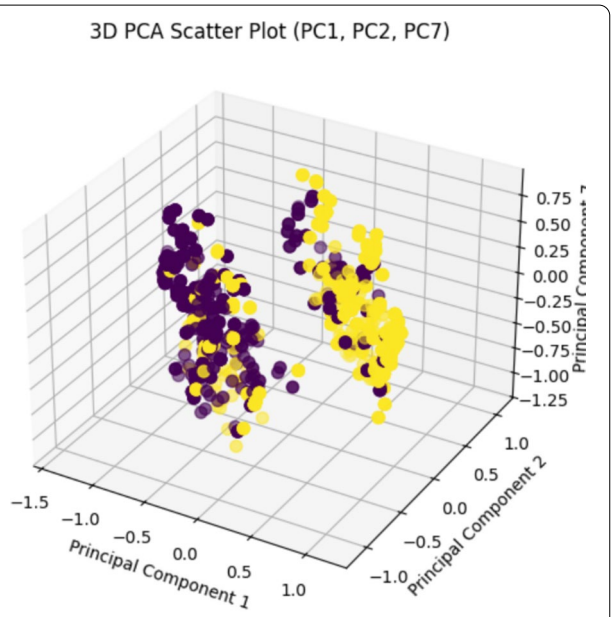

Fig. 2 PCA of LDF-dataset

revealed clearer and more compact class clusters, suggesting that the LDF transformation enhances discriminative structure by embedding uncertainty information within the feature space.

Hence, the central motivation of this study is to translate uncertainty into an actionable predictive signal by integrating LDFS with established supervised learners-Logistic Regression, Decision Tree, SVM, and XGBoost-for heart-disease prediction. This integration aims to

- capture the gradations of medical ambiguity,
- reduce preprocessing complexity by handling numerical and categorical data within a unified fuzzy framework, and
- enhance diagnostic recall and interpretability without compromising computational efficiency.

Structure of the study

The remainder of this paper is organized as follows. Section "Materials and methods" describes the dataset, preprocessing steps, and the mathematical foundation of the Linear Diophantine Fuzzy framework. Section "A hybridisation of LDF in ML approaches in heart disease prediction" outlines the integration of LDFS with supervised machine-learning models and the experimental design. Section "Analysis and Discussion" presents and analyzes the results, which also discusses comparative findings and implications for medical diagnosis. Finally, Section "Conclusion" concludes the study and highlights directions for future research.

Materials and methods

Dataset preprocessing

This study used the Kaggle heart disease prediction dataset. Before analysis, the data set was subjected to a comprehensive data cleaning and pre-processing process to ensure quality and reliability. These procedures addressed the issues related to missing and null values and were performed based on the following key criteria: completeness, validity, consistency, uniqueness, and precision.

1. **Completeness** Completeness ensures that our dataset contains all the necessary information required for analysis. There are no missing or null values identified to prevent bias and maintain the reliability of the data set.
2. **Validity** Validity checks confirm that data values lie within the predefined ranges. The attributes like age, blood pressure, and cholesterol levels are validated to ensure that the data represent realistic conditions.
3. **Consistency** Consistency ensures uniformity of data across the dataset. It involves verifying that related attributes do not contain conflicting information (e.g., ensuring that a "sex" attribute coded as 0 or 1 is consistently interpreted across all records).
4. **Uniqueness** There are no duplicate records found in this dataset. This step ensures the unique observation of each instance in the dataset.
5. **Accuracy** Finally, the dataset undergoes the accuracy verification, which ensures that the data exactly represent the real-world observation.

ML algorithms

In this study, several well-established supervised machine learning (ML) algorithms were employed to construct predictive models for heart disease diagnosis. These algorithms–Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost)–were selected for their proven effectiveness, interpretability, and complementary learning mechanisms. Together, they represent a comprehensive suite of linear, tree-based, and ensemble approaches, enabling a balanced evaluation of the proposed fuzzy-based framework.

Logistic regression

LR is a kind of method in supervised learning to solve classification problems. It gives the probability of an input falling in a certain class, which was in a binary setting. It is most suitable for the problems that have only two outcomes. LR is most widely applicable in all statistical models for binary classification. Even though it is termed as 'regression', it is a model for classification problems. This model mostly utilized the function called sigmoid to assign the instances with values between 0 and 1, which indicated the probability of being in a particular class. The primary advantage of LR is its accuracy and efficiency in smaller data, and also its interpretability.

Decision tree

A DT is a human-mimicking decision-making algorithm. IT divides the data step-by-step, like the branches in a tree, based on certain conditions to attain the conclusion. Each internal node consists of features, each branch symbolizes a decision rule, and each terminal node(leaf) indicates a class category or predicted value. The DT is generated by identifying the feature that separates the data best at each step. This algorithm often uses metrics like "gini impurity" and information gain. The main advantage of a DT is easy to visualize and suitable for understanding the decision processes. But it suffers from overfitting when the tree grows very deep.

Support vector machine

SVM typically identifies a clear boundary, known as a hyperplane, that differentiates the data of separate classes. It works by increasing the margin or distance between the two nearest points of different classes. It can effectively handle the models with both linearity and non-linearity, depending on the application of their different kernel functions, such as linear, polynomial, and RBF. It is well known for its higher generalisation ability in high-dimensional feature spaces. It is very effective in the dataset that has a higher number of features than observations.

XGBoost

Extreme Gradient Boosting(XGBoost) is a newfangled approach of gradient boosting that creates a typical DT in a sequential order. Each tree is trained to decrease the errors from the previous ensemble, and it will increase the performance of the model overall. The algorithm used gradient descent for optimizing the differential loss function. It regulates the parameters to control the complexity of the model and prevents overfitting. It is recognized for its efficiency in computation and accuracy in strong prediction. Its ability to handle missing values makes it highly suitable for structured data problems.

While these algorithms differ in mathematical formulation and model complexity, they all share a dependence on crisp numerical inputs. Consequently, their performance can deteriorate when facing data that contains linguistic, uncertain, or imprecise attributes–a frequent scenario in clinical datasets. To overcome these limitations, we introduce a novel LDF-based enhancement in the next section. This approach reformulates the dataset representation, allowing the models to process both categorical and numerical features within a unified fuzzy framework, thereby improving robustness, interpretability, and diagnostic reliability.

Fuzzification

Fuzzification refers to the formulation of membership functions of a fuzzy set. There are a limitless number of approaches to characterize fuzziness to depict the membership graphically, which describe fuzziness.

Definition 2.1 Let Δ be the universe of discourse. The FS over Δ is characterized as

$$F = \{(\mathfrak{k}, \mu(\mathfrak{k})) : \mathfrak{k} \in \Delta\}$$

where $\mu \in [0, 1]$ represents the membership degree of an element, that how much it belongs to the set.

Definition 2.2 Let Δ be the universe of discourse. The IFS over Δ is characterized as

$$I = \{(\mathfrak{k}, \langle \mu(\mathfrak{k}), \nu(\mathfrak{k}) \rangle) : \mathfrak{k} \in \Delta\}$$

where $\mu, \nu \in [0, 1]$ represent the membership degree and non-membership of an element belonging to the set.

The graphical representations of degree functions can be depicted in various shapes, such as triangles, trapezoids, or bell curves, as long as they accurately exemplify

the classification of information within the system. The straight lines are easy to represent the degree function. Straight lines are the easiest way to represent membership and non-membership functions.

$$\mu_{straight_line}(\mathfrak{k}) = \begin{cases} 0 & \mathfrak{k} \leq c \\ \frac{\mathfrak{k}-c}{d-c} & c < \mathfrak{k} < d \\ 1-\epsilon & \mathfrak{k} \geq d \end{cases}$$

Intuitive fuzzy triangular functions consist of three points creating a triangle, while intuitionistic fuzzy trapezoidal functions are simply a truncated triangle curve with a flat top.

$$\mu_{triangular}(\mathfrak{k}) = \begin{cases} 0 & \mathfrak{k} \leq c \\ \frac{\mathfrak{k}-c}{d-c} & c < \mathfrak{k} \leq d \\ \frac{e-\mathfrak{k}}{e-d} & d < \mathfrak{k} < e \\ 1-\epsilon & \mathfrak{k} \geq e \end{cases}$$

Here, the epsilon was chosen based on the level of uncertainty. Smooth curves are used to generate fuzzy Gaussian and bell-shaped functions.

$$\mu_{Gaussian}(\mathfrak{k}) = exp\left(\frac{-(\mathfrak{k}-m)^2}{2k^2}\right) - \epsilon$$

These functions are also simple curves that can be open to the left or right. Polynomial-based curves generate intuitive fuzzy S- and Z-shaped functions. Based on the meaning of membership degree, the function for fuzzification differs.

A hybridisation of LDF in ML approaches in heart disease prediction

This section proffers the comprehensive methodology of this study, encompassing the integration of LDFSs with various ML algorithms for heart disease prediction. Each phase of the methodology, ranging from dataset description, fuzzification, and model training to performance evaluation, is systematically elaborated in the following subsections. This step-by-step explanation ensures clarity in understanding the transformation and learning process. The general workflow of the introduced framework is exemplified in the Fig. 3, which outlines the key stages from data collection to performance comparison between crisp and fuzzified data sets.

Dataset description

The heart disease dataset, which was utilized in this research, was openly available on the website Kaggle [21]. It consists of 1888 patient records with 14 attributes, including the target label which indicates the presence or absence of heart disease in patients. There are 13 attributes considered for each patient in this dataset, both categorical and numerical. The description of 13 attributes is provided as follows:

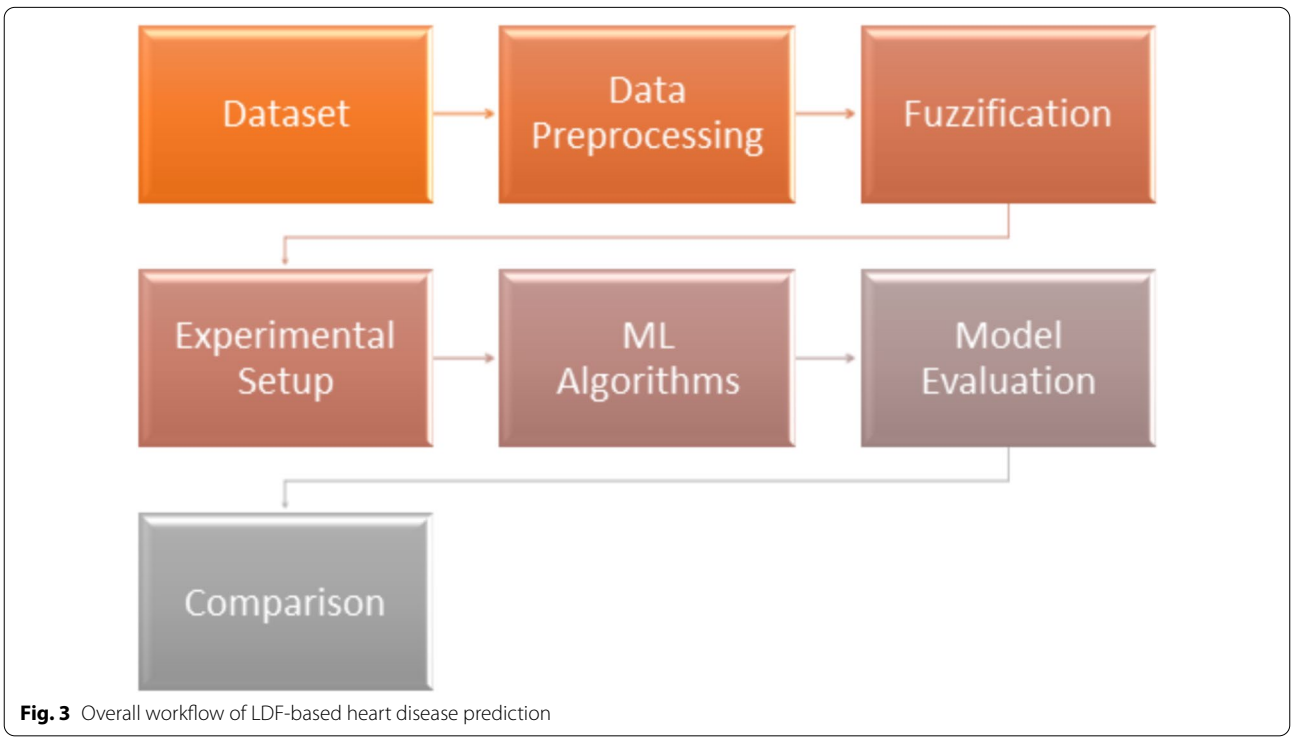

Fig. 3 Overall workflow of LDF-based heart disease prediction

- age - Indicates the age of the patient in years.
- sex - Indicates the gender of the patient. 1 represents male and 0 represents female.
- cp - It describes the type of chest pain experienced. Here, there are four types of chest pain categorized.
- trestbps - It represents the resting blood pressure, which is measured in mm/Hg.
- chol - It indicates the level of serum cholesterol, which was measured in mg/dl units.
- fbs - It describes the fasting blood sugar level. If the level is more than 120 mg/dl, the data shows 1; otherwise, 0.
- restecg - It shows the results of the resting electrocardiogram classified in 4 levels.
- thalachh - It represents the maximum heart rate attained during exercise.
- exang - It indicates whether angina was induced by exercise (1 indicates yes and 0 represents no).
- oldpeak - It represents ST depression induced by exercise relative to rest.
- slope - It describes the slope of the peak.
- ca - Number of major coronary vessels (0-4) colored by fluoroscopy.
- thal (thalassemia) - It represents a type of blood disorder
- target - Output variable showing the presence or absence of heart disease.

Among these, the attributes such as age, trestbps, chol, thalachh, and old peak are numerical features. Other than these are categorical features. The target indicates the likelihood of heart disease, where 1 represents the presence of heart disease and 0 represents the absence of heart disease.

For conventional ML algorithms, the categorical features must be converted into numerical form using encoding methods. However, in the current study, this type of conversion was not required. Because our model employs a fuzzification process that transforms both numerical and categorical features into fuzzy values using membership functions. Thus, it reduces the information loss and computational complexity. It preserves the interpretability and semantic meaning of the data. The range of each feature and the bivariate analysis of each feature with the target are shown in the Figs. 4 and 5.

Linear Diophantine fuzzification

The term LD-fuzzification functions relates to the construction of membership and non-membership functions along with their reference parameters for an LDFS.

Definition 3.1 Let Δ be the universe of discourse. The LDFS over Δ is symbolised as

$$\mathfrak{L} = \{\mathfrak{k}, \langle \mu(\mathfrak{k}), \nu(\mathfrak{k})\rangle, \langle \alpha(\mathfrak{k}), \beta(\mathfrak{k})\rangle : \mathfrak{k} \in \Delta\}$$

where $\mu(\mathfrak{k})$, $\nu(\mathfrak{k})$, $\alpha(\mathfrak{k})$, $\beta(\mathfrak{k}) \in [0,1]$ represents the membership, non-membership, reference parameters respective to membership and non-membership respectively.

Since the LDFS framework contains four degree values, we have to define four corresponding functions for each attribute. Based on the observed ranges, we construct straight-line functions using the max-min of the range for both the membership and non-membership degrees to reduce the complexity of fuzzification and maintain interpretability. Let us consider the feature 'chol'. As demonstrated in Fig. 5, the cholesterol values in the dataset range from 128 to 564. Based on this range, the membership and non-membership functions are constructed.

$$\mu_{chol}(\mathfrak{k}) = \begin{cases} 0 & \mathfrak{k} \leq 128 \\ \frac{\mathfrak{k}-128}{564-128} & 128 < \mathfrak{k} < 564 \\ 1-0.01 & \mathfrak{k} \geq 564 \end{cases}$$

$$\nu_{chol}(\mathfrak{k}) = \begin{cases} 0.99 & \mathfrak{k} \leq 128 \\ 1-\frac{\mathfrak{k}-128}{564-128} & 128 < \mathfrak{k} < 564 \\ 0 & \mathfrak{k} \geq 564 \end{cases}$$

The reference parameters for constructing these functions are derived from physicians' opinions with medical insights. Practically, different medical experts may exhibit varying cholesterol thresholds associated with elevated risk levels. However, in this study, we generalize the range 120-600 mg/dL as a high-risk interval for heart disease.

$$\alpha_{chol}(\mathfrak{k}) = \begin{cases} 0 & \mathfrak{k} \leq 120 \\ \frac{\mathfrak{k}-120}{600-120} & 120 < \mathfrak{k} < 600 \\ 1-0.01 & \mathfrak{k} \geq 600 \end{cases}$$

$$\beta_{chol}(\mathfrak{k}) = \begin{cases} 0.99 & \mathfrak{k} \leq 120 \\ 1-\frac{\mathfrak{k}-120}{600-120} & 120 < \mathfrak{k} < 600 \\ 0 & \mathfrak{k} \geq 600 \end{cases}$$

For example, A patient with a cholesterol level of 200 mg/dL, for instance, is in the clinically moderate range. The membership and non-membership degrees in the LDF representation show that this value is neither obviously average nor extremely high. The reference parameters also measure the degree to which this cholesterol level influences the probability of heart disease as opposed to its absence. This distinction highlights the possibility that a moderate cholesterol level may have an ambiguous or contextually dependent impact and may also be linked to other disorders. Instead of rendering a deterministic decision, the LDF paradigm offers interpretable insight

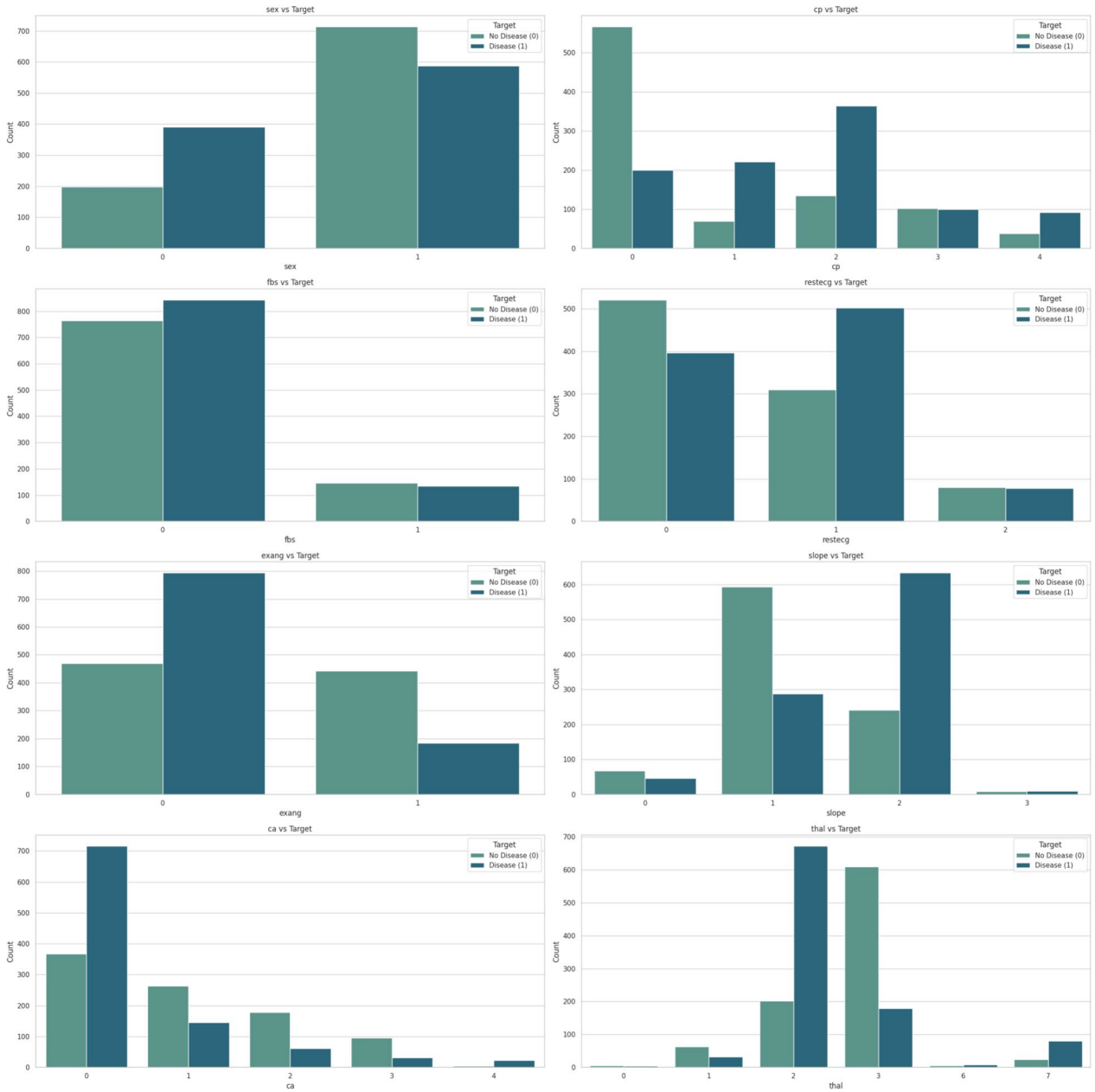

Fig. 4 Categorical Features vs Target

into how cholesterol influences diagnostic uncertainty by explicitly modeling this overlap. A four-fold informational framework of LDFS, which permits more nuanced representation of clinical evidence, is produced by these reference parameters, which offer contextual scaling and semantic grounding of ambiguity. In diagnostic contexts, where finer uncertainty refinement might enable more informed clinical decision-making, this additional contextual information is highly beneficial.

For all 13 features, linear membership and non-membership functions are constructed using dataset-derived minimum and maximum values, while the reference parameters are fixed based on clinically defined threshold ranges to contextualize uncertainty.

Experimental setup

In this study, the heart disease prediction dataset, comprising 1,888 instances and 14 attributes, is utilized.

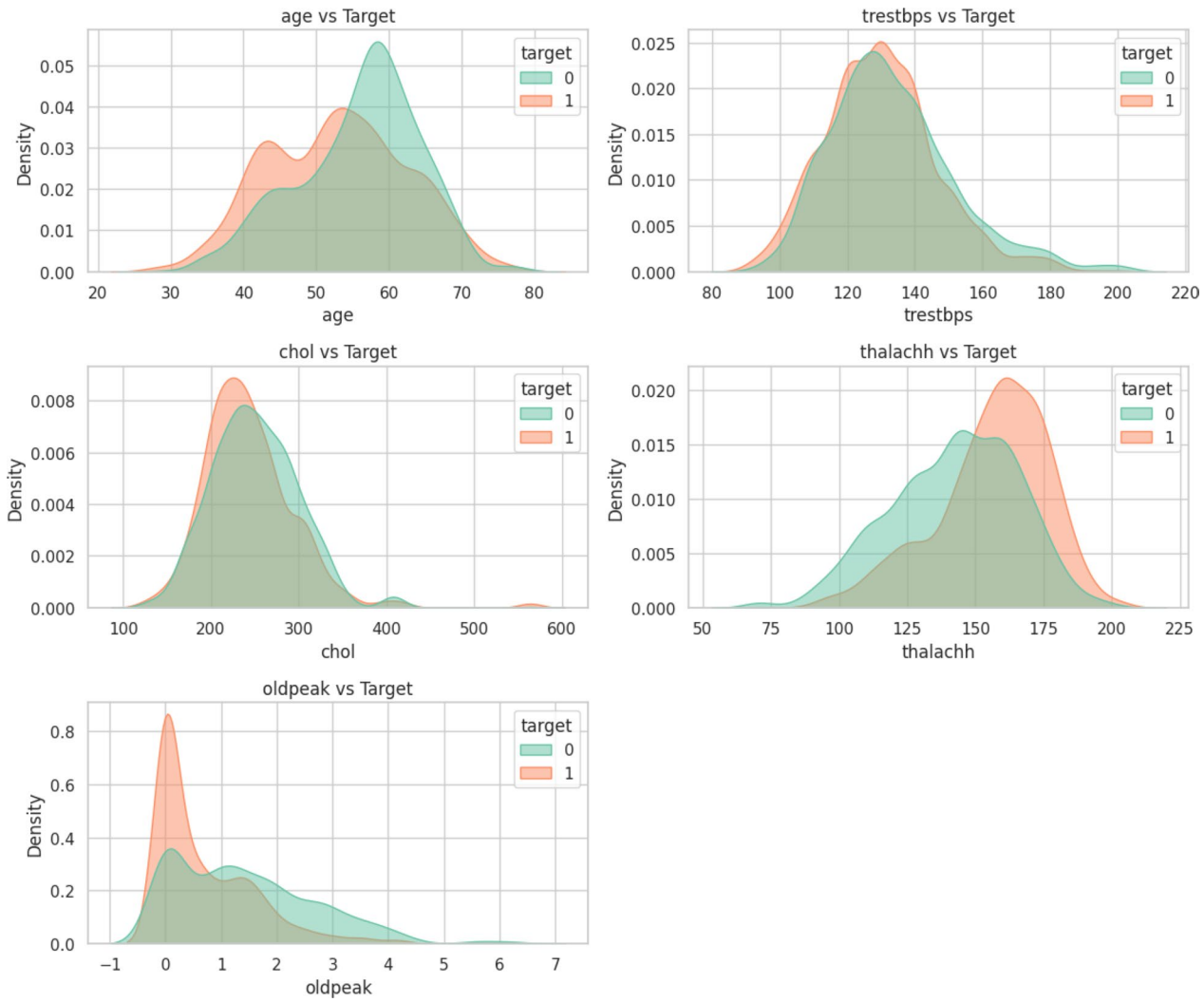

Fig. 5 Numerical Features vs Target

The data undergoes all necessary preprocessing checks before the training. Also, the crisp dataset was converted into an LDF dataset to capture the uncertainty in medical data through a fuzzification process. All the algorithms were trained and tested on the Google Colab platform, which enables a cloud-based environment with GPU acceleration for making the computations effectively. The implementation was done on Python 3.10 with libraries like NumPy, Pandas, Scikit-Learn, Scikit-fuzzy, and Matplotlib.

In this study, ML algorithms like LR, DT, SVM, and XGB were implemented to evaluate the performance. The performance of the initiated approach is evaluated and compared with previous models in terms of evaluation metrics.

Preprocessing

For a consistent class distribution, the dataset was divided into training and testing sets using an 80:20 stratified sampling technique. Using stratified k-fold cross-validation on the training set, model selection and hyperparameter tuning were carried out. The usual one-hot encoding method was used to encode categorical characteristics for the crisp dataset, and then normalization was applied. LDF-ML Model does not require any encoding for categorical features due to fuzzification. However, the same normalization process was used to guarantee comparability. For both crisp and LDF-based models, the same model architectures and search regions were used for hyperparameter tuning. To prevent information leakage, all preprocessing and tuning procedures

Table 1 Comparison of LDF-logistic regression and crisp logistic regression model

Performance Metrics	LDF-LR	LR
Accuracy	**0.8333**	0.8148
Precision	**0.8182**	0.7971
Recall	**0.8724**	0.8622
F1-score	**0.8444**	0.8284

Table 2 Comparison of LDF-decision tree and crisp decision tree model

Performance Metrics	LDF-DT	DT
Accuracy	**0.8651**	0.8465
Precision	**0.868**	0.8415
Recall	**0.8724**	0.8673
F1-score	**0.8702**	0.8542

Table 3 Comparison of LDF-support vendor machine and crisp support vector machine model

Performance Metrics	LDF-SVM	SVM
Accuracy	0.8889	**0.9179**
Precision	0.8667	**0.9319**
Recall	**0.9286**	0.9081
F1-score	0.8966	**0.9199**

Table 4 Comparison of LDF-XGBoost and Crisp XGBoost model

Performance Metrics	LDF-XGB	XGB
Accuracy	**0.9735**	0.9470
Precision	**0.9559**	0.9356
Recall	**0.9949**	0.9642
F1	**0.975**	0.9497

were performed exclusively within the training folds. To maintain class imbalance, stratified cross-validation was used in every trial. To prevent information leaking, LDF fuzzification parameters for each fold were calculated using just training data and then applied to the test set.

Experimental results

Table 1 demonstrated the performance of the LR model on both the original crisp dataset and the LDF-dataset. The results clearly demonstrate a noticeable improvement in all evaluation metrics when the model is trained using the LDF-dataset. The LDF-LR model gains additional information from the reference parameters, which helps to distinguish it more effectively and makes it easier to handle uncertain information. This shows that even a simple linear model like LR can achieve better decisions when integrated with the proposed linear diophantine fuzzification.

Table 2 summarizes the results from the DT algorithm before and after the linear diophantine fuzzification process. The LDF-DT model shows a consistent increase in performance metrics compared to the conventional DT model. The ensemble nature of the DT algorithm becomes more effective when combined with LDF inputs, as the uncertainty information embedded in each feature allows individual trees to make more balanced decisions. This improvement highlights that the integration of linear diophantine fuzzification enables the capture of complex feature interactions and uncertainty-driven patterns more efficiently.

Table 3 presents the comparative analysis between the SVM model with LDF-based SVM. In this case, the performance gain is increased more efficiently in all metrics. When the SVM model is trained on the LDF dataset, it benefits from uncertainty-based representation of the data, which helps the kernel function to identify more flexible and accurate decisions. This is particularly useful in our medical diagnosis datasets, where overlapping and ambiguous data samples often occur. The enhanced performance of LDF-SVM represents that incorporating linear diophantine fuzzification can significantly reduce misclassification errors and improve the reliability of predictions in uncertain environments.

Table 4 illustrates the comparative results between the XGB classifier trained on the original crisp dataset and its counterpart, LDF-XGB, trained on the LDF-based dataset. The LDF-XGB model achieves consistently higher evaluation metrics, reflecting the ability of the LDF integration to enhance the capacity of decision-making in models. This incorporation of LDF values allows the algorithm to effectively handle uncertain and imprecise data samples by adjusting decision weights. This result represents the stronger generalization of performance metrics, and it improves the stability of the data. This shows that the proposed LDF integration is appropriate for advanced ensemble models.

Thus, across all comparisons, the introduction of the LDF-based dataset significantly enhances the performance of traditional ML models. The consistent improvement across various algorithms such as LR, DT, SVM, and XGB shows the robustness of the proposed LDF-ML techniques. These results confirm that assimilating LDF value principles can be a promising strategy for improving predictive accuracy in medical diagnosis.

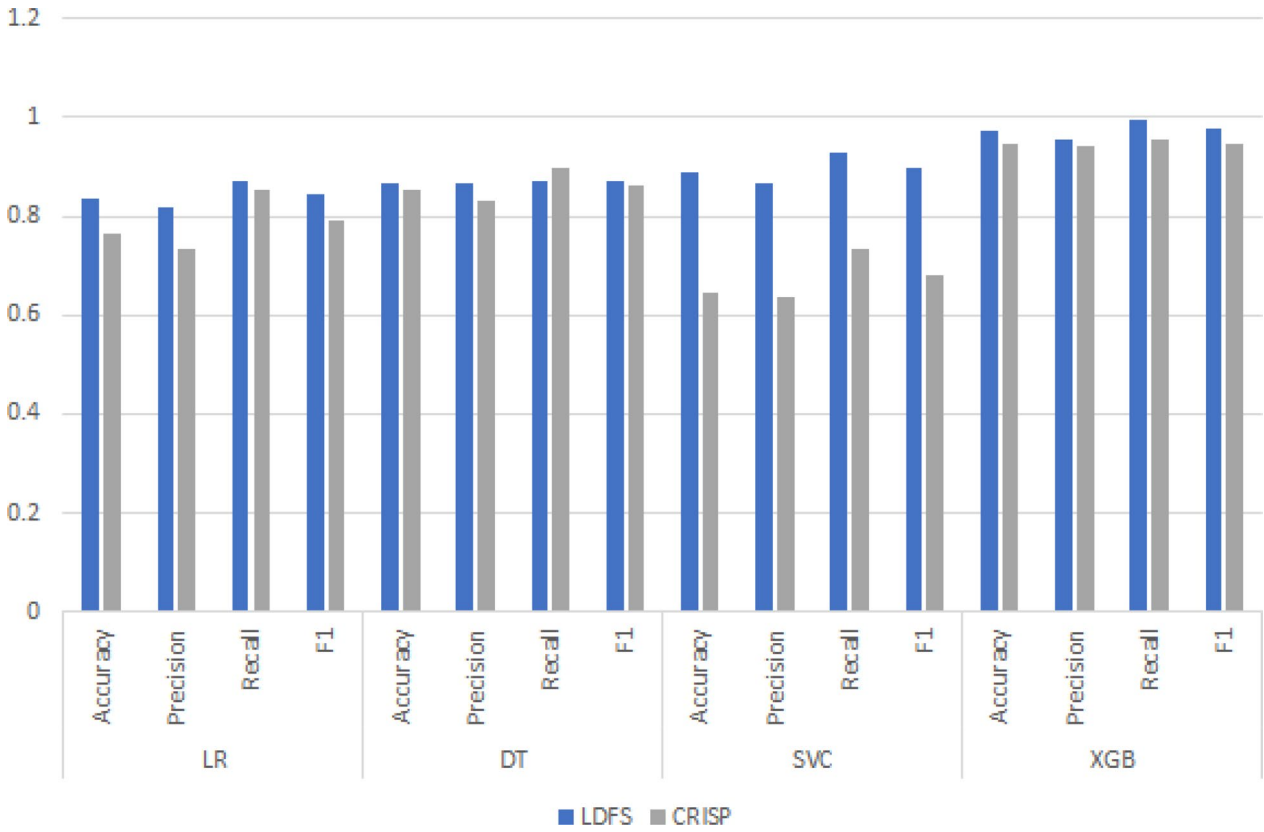

Fig. 6 Comparison of various ML algorithms in LDF and Original Crisp Dataset

Figure 6 illustrates the overall comparisons of four algorithms across all performance metrics.

Discusssion

Thus, the proposed LDF-ML models demonstrate superior performance compared to conventional ML models. The inclusion of fuzzification provides more efficient results in heart disease prediction. Among all evaluated models, LDF-XGB exhibits the highest values in all evaluation metrics. In particular, the recall value of LDF-XGB gives 0.9949, which is more significant for medical diagnosis. Since our dataset is for heart disease prediction, it is crucial to identify positive cases correctly as well as to minimize the false negatives. In that way, we could avoid medications for unaffected individuals. Hence, the recall is an important metric in such medical diagnosis. Thus, the higher value of recall in LDF-XGB confirms the robustness of our proposed method. As a result, the LDF-XGB model can be represented as the most appropriate and effective method for the specified dataset in terms of both numerical outcomes and clinical significance. The Fig. 7 represents the ROC curve for the original dataset and Fig. 8 for the LDF-Based dataset. It clearly demonstrates that the ROC of LDF-based models has slightly higher AUC values for LR and DT, which confirms the better classification performance. Thus, overall, our proposed LDF-based ML models are superior to the conventional ML models.

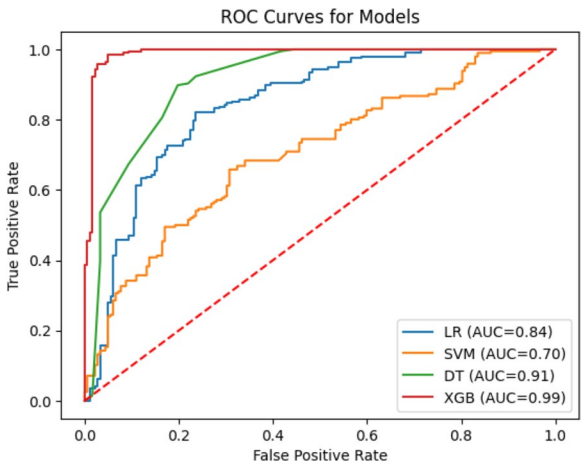

Fig. 7 ROC curve for original crisp dataset

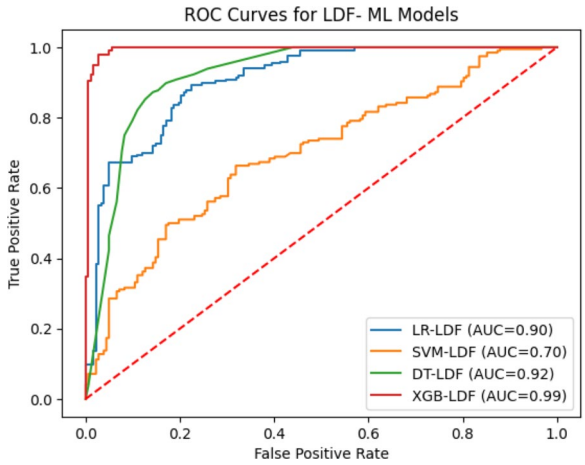

Fig. 8 ROC curve for LDF-based dataset

Analysis and discussion

Comparative analysis

This section explores the comprehensive comparison between the datasets generated using the Linear Diophantine Fuzzification, the Intuitionistic Fuzzification, and the normal fuzzification process representations. FS and IFS transformations were implemented using the same linear membership functions and feature bounds as LDFS for a fair comparison; the only difference was the number of uncertainty components modeled. The experimental results clearly demonstrate that the LDF-based dataset produces better outcomes regarding the performance evaluation. While the DT algorithm produces comparable results across all three datasets, several others exhibit a better improvement on the LDF-based dataset. This shows the enhancement of LDF-ML models over traditional fuzzy and intuitionistic fuzzy frameworks. Thus, the LDF dataset exhibits the highest suitability in the enhancement of accurate prediction. The associated Table 5 summarizes the comparative values.

Statistical analysis

The performance between LDF and Crisp ML was analysed by conducting the statistical analysis (Table 6). The Shapiro test was initially used to check the normality of the performance distributions. Based on the normality results, either a paired t-test (for normally distributed data) or the Wilcoxon signed-rank test (for non-normal data) was applied to compare the two approaches. Statistical significance was determined using the resulting p-values, and in every instance, $p > 0.05$ indicated no statistically significant difference across the assessed metrics. These results suggest that the proposed approach provides an improved representation of uncertainty while achieving predictive performance on par with traditional models. Uncertainty-aware modeling is essential for risk assessment and well-informed diagnosis in clinical decision-making, especially in circumstances that are unclear or borderline. Instead of being a rival to crisp models, the LDF framework offers a clinically useful substitute by explicitly capturing such uncertainty without sacrificing predictive accuracy.

Table 5 Comparison of LR, DT, SVM, XGB over LDFS, IFS and FS generalised dataset

Algorithms	Metrics	LDFS	IFS	FS
LR	Accuracy	**0.8333**	0.8307	0.8307
	Precision	0.8182	**0.8267**	0.8204
	Recall	**0.8724**	0.852	0.8622
	F1	**0.8444**	0.8392	0.8408
DT	Accuracy	**0.8651**	**0.8651**	**0.8651**
	Precision	**0.868**	**0.868**	**0.868**
	Recall	**0.8724**	**0.8724**	**0.8724**
	F1	**0.8702**	**0.8702**	**0.8702**
SVM	Accuracy	0.8889	0.8862	**0.8995**
	Precision	0.8667	0.8732	**0.899**
	Recall	**0.9286**	0.9133	0.9082
	F1	0.8966	0.8928	**0.9036**
XGB	Accuracy	**0.9735**	0.955	0.9524
	Precision	**0.9559**	0.9366	0.9363
	Recall	**0.9949**	0.9796	0.9745
	F1	**0.975**	0.9576	0.955

Table 6 Statistical analysis: crisp dataset Vs LDF dataset

Algorithm	Metrics	Shapiro p-value	Wilcoxon/ Paired t-test	Difference
LR	AUC	0.216	0.236	No statistical difference
	F1	0.722	0.089	No statistical difference
DT	AUC	0.397	0.998	No statistical difference
	F1	0.202	0.643	No statistical difference
SVC	AUC	0.01	1.705	No statistical difference
	F1	0.011	1.934	No statistical difference
XGB	AUC	0.002	0.482	No statistical difference
	F1	0.007	0.668	No statistical difference

Merits of this study

The proposed framework highlights several significant advantages that contribute to advancing medical diagnostic accuracy:

- **Enhancement in Uncertainty Handling** The incorporation of the Linear Diophantine Fuzzy (LDF) framework with ML algorithms effectively captures and models uncertainty and vagueness.Compared to current methods like IFS, PFS, NS, or type-2 fuzzy models, LDFS employs four distinct degrees to represent each attribute. This information structure enabled more detailed modeling of uncertainty and feature influence in medical tabular data.
- **Improved Classification Accuracy** The comparative analysis demonstrates that LDF-based ML models consistently yield higher performance over traditional crisp, fuzzy, and intuitionistic fuzzy datasets in terms of accuracy, precision, and recall.
- **No Need for Categorical Encoding** Conventional ML algorithms require preprocessing pipelines for encoding categorical and numerical data, but the LDF-based ML algorithms inherently manage both types through fuzzification, and it simplifies the preprocessing step.
- **Applicability in Uncertain Environments** The LDF framework enables the application of ML models in real-world uncertain environments, such as medical diagnosis and risk prediction.
- **Versatility Across Algorithms** The LDF dataset enhances the performance of various ML algorithms like LR, DT, SVM, and XGB.
- **Foundation for Future Research** The findings encourage working on the integration of LDF concepts with advanced deep learning techniques, dimensionality reduction, and neural networks to enhance performance in diagnosis.

Conclusion

This research presents a comprehensive investigation of the integration of LDF in ML algorithms for heart disease diagnosis. The crisp dataset was transformed into an LDF-based fuzzy representation using membership functions. Thus, the proposed method effectively captured the uncertainty and vagueness in medical data. The study introduced the hybrid LDF-based ML-algorithms like LDF-LR, LDF-DT, LDF-SVM, and LDF-XGB. The evaluation was performed on both crisp and LDF-based datasets. Also, compared the results on ML algorithms. The result confirms the excellence of the proposed hybrid method by its high performance across various metrics in the taken dataset. The improvement in metrics confirms the strength and durability of the advanced LDF-ML framework. The study highlights the potential of integrating LDF systems with ML techniques to achieve superior classification outcomes. The findings underscore the significance of leveraging hybrid ML-LDF models for heart disease diagnosis, and this leads to new ways for improving healthcare diagnosis. The future work will be extended on a neural network model integrated with LDF sets to explore deep fuzzy structures.

Abbreviations

ML: Machine learning; LR: Logistic regression; DT: Decision tree; SVM: Support vector machine; XGB: Extreme gradient boosting; FS: Fuzzy set; IFS: Intuitionistic fuzzy set; LDFS: Linear Diophantine fuzzy set.

Data availability

The dataset analysed during this study is publicly accessible via Kaggle at: https://www.kaggle.com/datasets/mfarhaannazirkhan/heart-dataset. No further permissions were required for access. Any processed version of the data used in the study is available from the corresponding author upon reasonable request.

Author details

[1]Center of Computational Biology, SRM Institute of Science and Technology, Ramapuram, Chennai, India. [2]Center for Research, Easwari Engineering College, Chennai, India. [3]Department of Mathematics, Alagappa University, Karaikudi, India. [4]Department of Mathematics, Faculty of Arts and Science, 10145 Balikesir, Turkey. [5]Department of Mathematical Sciences and Philosophy, University of Illinois Springfield, Springfield, IL, USA. [6]Department of Mathematics, Yildiz Technical University, Faculty of Arts and Science, 34220 Esenler, Istanbul, Türkiye.

Received: 11 November 2025 Accepted: 8 February 2026
Published online: 24 February 2026

References

1. Zhou W, Liu X, Bai H, He L. Intelligent medical diagnosis and treatment for diabetes with deep convolutional fuzzy neural networks. Inf Sci. 2024;677:120802. https://doi.org/10.1016/j.ins.2024.120802.
2. Zhang T, Xue G. Fuzzy attention-based deep neural networks for acute lymphoblastic leukemia diagnosis. Appl Soft Comput. 2025;171:112810. https://doi.org/10.1016/j.asoc.2025.112810.
3. Vyas S, Gupta S, Bhargava D, Boddu R. Fuzzy logic system implementation on the performance parameters of health data management frameworks. J Healthc Eng. 2022;2022:9382322.
4. Dehghani Saryazdi M, Mostafaeipour A. Identification and validation of key predictive factors for heart attack diagnosis using machine learning and fuzzy clustering. Eng Appl Artif Intell. 2025;142:109968. https://doi.org/10.1016/j.engappai.2024.109968.
5. Khushal R, Fatima U. Fuzzy quantum machine learning logic for optimized disease prediction. Comput Biol Med. 2025;192:110315. https://doi.org/10.1016/j.compbiomed.2025.110315.
6. Zadeh LA. Fuzzy sets. Inf Control. 1965;8:338–53. https://doi.org/10.1016/S0019-9958(65)90241-X.
7. Nanehkaran YA, et al. Anomaly detection in heart disease using a density-based unsupervised approach. Wirel Commun Mob Comput. 2022;2022:6913043.
8. Liu Q, Hu W, Yang K, Yang J. Risk assessment of urban underground logistics system operations in built-up areas using a hybrid fuzzy Bayesian network and machine learning approach. Comput Ind Eng. 2025;207:111295. https://doi.org/10.1016/j.cie.2025.111295.
9. Jayalakshmi M, et al. Fuzzy logic-based health monitoring system for COVID-19 patients. Comput Mater Contin. 2021;67:2431–47.

10. Reddy GT, Reddy MPK, Lakshmanna K, Rajput DS, Kaluri R, Srivastava G. Hybrid genetic algorithm and fuzzy logic classifier for heart disease diagnosis. Evol Intell. 2020;13:185–96. https://doi.org/10.1007/s12065-019-00327-1.
11. Lohani QMD, Solanki R, Muhuri PK. A convergence theorem and an experimental study of intuitionistic fuzzy c-mean algorithm over machine learning dataset. Appl Soft Comput. 2018;71:1176–88. https://doi.org/10.1016/j.asoc.2018.04.014.
12. Atanassov KT. Intuitionistic fuzzy sets. Fuzzy Sets Syst. 1986;20:87–96. https://doi.org/10.1016/S0165-0114(86)80034-3.
13. Khan VA, Yadav AK, Arshad M, Akhtar N. Lung cancer prediction using an enhanced neutrosophic set combined with a machine learning approach. Neutrosophic Sets Syst. 2025;88:1.
14. Zhou T, Wang H, Geng S, Ju H, Huang J, Fu F, et al. F2CAU-Net: a dual fuzzy medical image segmentation cascade method based on fuzzy feature learning. Appl Soft Comput. 2025;184:113692. https://doi.org/10.1016/j.asoc.2025.113692.
15. Bai L, Chen X, Wang Z, Shao Y-H. Safe intuitionistic fuzzy twin support vector machine for semi-supervised learning. Appl Soft Comput. 2022;123:108906. https://doi.org/10.1016/j.asoc.2022.108906.
16. Riaz M, Hashmi MR. Linear Diophantine fuzzy set and its applications towards multi-attribute decision-making problems. J Intell Fuzzy Syst. 2019;37:5417–39.
17. Kannan J, Jayakumar V, Saeed M, Alballa T, Khalifa HAE-W, Gomaa HG. Linear Diophantine fuzzy clustering algorithm based on correlation coefficient with application to logistic efficiency. IEEE Access. 2024. https://doi.org/10.1109/ACCESS.2024.3371986.
18. Jayakumar V, Kannan J, Kausar N, Deveci M, Wen X. Multicriteria group decision making for prioritizing IoT risk factors using linear Diophantine fuzzy sets and MARCOS method. Granul Comput. 2024;9:56. https://doi.org/10.1007/s41066-024-00480-8.
19. Kannan J, Jayakumar V, Kausar N, Pamucar D, Simic V. Enhancing decision-making with linear Diophantine multi-fuzzy set using novel information measures. Sci Rep. 2024;14:79725. https://doi.org/10.1038/s41598-024-79725-0.
20. Vimala J, Garg H, Jeevitha K. Prognostication of myocardial infarction using lattice ordered linear Diophantine multi-fuzzy soft set. Int J Fuzzy Syst. 2024;26:44–59. https://doi.org/10.1007/s40815-023-01574-2.
21. Nazirkhan F. Heart disease prediction dataset. Kaggle Dataset (2024). https://www.kaggle.com/datasets/mfarhaannazirkhan/heart-dataset.