Article:

Miller, S.T., Logan, K.A., Anderson, R. et al. (2026) Multi-perspective machine learning MPML: a high-performance and interpretable ensemble method for heart disease prediction. Machine Learning with Applications, 23. 100836. ISSN: 2666-8270

https://doi.org/10.1016/j.mlwa.2026.100836

Multi-Perspective Machine Learning MPML: A High-Performance and Interpretable Ensemble Method for Heart Disease Prediction

Sean T Miller[a], Keaton A Logan[b], Ricardo Anderson[a], Patricia E Cowell[c],
Curtis Busby-Earle[a], Lisa-Dionne Morris[d]

[a]*The University of the West Indies Mona sean.miller@uwi.edu,*
ricardo.anderson@uwi.edu,
curtis.busbyearle@uwi.edu, Kingston, Jamaica
[b]*Caribbean Maritime University klogan@faculty.cmu.edu.jm, Kingston, Jamaica*
[c]*The University of Sheffield p.e.cowell@sheffield.ac.uk, Sheffield, United Kingdom*
[d]*The University of Leeds L.D.Morris@leeds.ac.uk, Leeds, United Kingdom*

Abstract

Machine Learning (ML) has demonstrated strong predictive capabilities in healthcare, often surpassing human performance in pattern recognition and decision-making. However, many high-performing models lack interpretability, which is critical in clinical settings where understanding and trusting predictions is essential. To achieve our objective, we proposed a Multi-Perspective machine learning framework (MPML) that combines established base classifiers with structured perspective-based design and interpretability pipeline. MPML organises features into meaningful subsets, or perspectives, enabling both global and instance-level interpretability. Unlike traditional ensemble methods such as Bagging, Boosting, and Random Forest, MPML delivers significantly higher-quality predictions across all evaluation metrics while maintaining a transparent structure. Applied to a heart disease dataset, MPML not only improves predictive accuracy but also provides detailed, accessible explanations for individual patient outcomes, advancing the potential for practical and ethical deployment of ML in healthcare.

Keywords:
Machine Learning, Healthcare, Explainable AI, Predictions, Algorithmic Accountability

1. Introduction

Machine Learning (ML) has become a powerful tool in data-driven domains such as healthcare, where accurate predictions and informed decision-making are critical. However, many high-performing ML models function as "black boxes," offering little transparency into how predictions are made (Rudin, 2019). This lack of interpretability poses significant challenges in domains where trust, accountability, and ethical considerations are paramount. To address this gap, we propose Multi-Perspective Machine Learning (MPML). This ensemble approach integrates multiple established techniques to achieve both high predictive performance and model interpretability.

Recent efforts to enhance machine learning in healthcare have increasingly focused on balancing predictive performance with interpretability, a challenge that traditional ensemble methods often fail to address. For example, one author (Topuz et al., 2025) emphasized the gap between highly accurate but opaque ensemble models and the need for interpretable AI in critical healthcare tasks. To address this, various hybrid frameworks have been proposed. Another study (Al-bakri et al., 2025) introduced a meta-learning-based ensemble for Alzheimer's diagnosis, combining predictive strength with transparent decision pathways. Work done in another study (Awe et al., 2025) demonstrated the use of LIME within ensemble models for malaria diagnosis, enhancing clinician trust in model outputs. Similarly, another group of researchers (Acharya et al., 2025) developed a stacking-based XAI approach for diabetes classification, improving interpretability without sacrificing accuracy. In contrast to these approaches, MPML provides a principled integration of multiple perspectives (feature groups formed from statistical correlations and expert knowledge) yielding not only higher predictive power but also inherently interpretable model behaviour. This allows domain experts to trace predictions back to relevant features and perspectives, aligning machine learning outputs with clinical reasoning.

MPML draws on the principles of multi-view learning (Zhao et al., 2017), which treats datasets as having multiple distinct yet complementary perspectives. By using feature selection and domain-informed subgrouping, MPML organizes features into meaningful subsets, or perspectives. These perspectives form the structural foundation of the ensemble, helping to capture diverse aspects of the data and improve predictive accuracy across standard ML metrics.

To support interpretability, MPML's architecture enables the isolation of feature groups and their individual contributions to predictions. This design is inspired by methods such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) (Panda & Mahanta, 2023). LIME generates perturbed samples around a given instance and analyses the resulting changes in the model's output and SHAP is a method that assigns importance scores to input features. By adapting these ideas within a structured ensemble, MPML provides interpretable outputs both at the instance level and across the model.

While MPML offers notable advantages in accuracy and transparency, these come with trade-offs. The added complexity of perspective construction and interpretability analysis introduces computational overhead and increases training time. Despite these limitations, MPML represents a promising step toward creating machine learning systems that are both effective and explainable, especially in sensitive, high-stakes environments like healthcare.

In this work, we make three main contributions. First, we introduce and formalize multi-perspective machine learning (MPML), a framework that uses domain knowledge to group features into clinically meaningful perspectives, each modelled by its own base learner. Second, we extend perturbation-based explanation methods to generate consistent feature-level and perspective-level impact scores for both local (per-patient) and global model behavior across all models in the stack. Third, we demonstrate MPML on both a small multi-source heart-disease dataset and a large cardiovascular dataset, showing that it can match or outperform strong ensemble baselines while providing interpretable insights into how each perspective contributes to the model's predictions.

The remainder of this paper is structured as follows: Section 2 - Related Work reviews prior studies relevant to our research. Section 3 - Multi-Perspective Machine Learning introduces the proposed MPML approach in detail. Section 4 - Datasets outlines the datasets used in our experiments. Section 5 - Experiments and Results presents the performance of MPML and other ensemble methods on the datasets, along with interpretability analyses using MPML. Section 6 - Discussion and Limitations compares the interpretive findings to established research in heart disease diagnosis and addresses the limitations of the study. Finally, Section 7 - Conclusion and Future Work summarizes the study's contributions and outlines potential directions for future research.

2. Related Work

Explaining machine learning models has become a critical area of research, particularly in domains where model transparency and accountability are essential, such as healthcare. A wide range of approaches have been proposed to make complex machine learning models more interpretable, addressing concerns over their “black-box” nature. Prominent examples include LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), both of which provide ways to approximate and understand how models arrive at specific predictions. These techniques have made significant contributions to improving interpretability, particularly for individual predictions in otherwise opaque models. In addition to interpretability techniques, ensemble learning methods such as Bagging and Boosting have become foundational in building robust and accurate predictive models. However, ensemble methods often increase model complexity, further exacerbating the challenge of interpretability (Bassan et al., 2025).

Beyond technical considerations, there is a growing body of work emphasizing the urgent need for explainable AI (XAI) in high-stakes domains like healthcare, where decisions can directly impact patient outcomes and where regulatory standards increasingly demand transparent and trustworthy models (Rudin, 2019). This work introduces MPML, which draws from both interpretability research and ensemble learning to address these challenges. MPML leverages ensemble principles to enhance performance while incorporating interpretability directly into the model's structure by design, rather than as an afterthought. As this section will discuss, MPML is specifically poised to meet the interpretability needs of healthcare applications by providing both global and local explanations, offering feature-level insights, and ensuring decision-making processes remain transparent without compromising predictive accuracy. We review existing work on ensemble methods, interpretability approaches, and explainable AI in healthcare, highlighting where MPML builds upon, diverges from, and advances these prior efforts.

Recent work has explored ensemble-based methods for disease prediction, including coronary heart disease classification using machine-learning ensembles (Gulati, Guleria, & Goyal, 2022), ensemble methods for non-invasive coronary-artery disease detection (Sapra, Sandhu, & Goyal, 2021) and stacking-based ensembles for infectious disease prediction (Mahajan et al., 2022). These methods typically focus on improving predictive performance and may use feature-importance or SHAP-style explanations at the feature level. In contrast, MPML organises predictors into clinically motivated perspectives and provides explanations at both feature and perspective levels, offering a structured view of how different clinical domains contribute to risk.

1. LIME (Local Interpretable Model-agnostic Explanations)

LIME is an Explainable AI (XAI) approach designed to provide interpretability for black-box models by locally approximating the model's behaviour around a specific prediction (Ribeiro et al., 2016). LIME works by generating perturbations of the input data and then analysing how these changes impact the model's predictions (Salih et al., 2024). It builds a simple, interpretable model (like a linear regression) to approximate the predictions of a more complex, opaque model for a particular instance. This local model allows users to gain insights into why the black-box model made a specific decision (Hassija et al., 2024).

One of the key advantages of LIME is that it is model-agnostic, meaning it can work with any machine learning model, regardless of the underlying architecture, whether it be neural networks, decision trees, or any other type of model (Ribeiro et al., 2016). Another strength of LIME is its ability to provide instance-level explanations. It helps users understand how each feature contributes to a specific prediction.

However, LIME has some limitations. Its primary focus is on providing local approximations of the model's behaviour around a specific instance, which means it doesn't offer a global view of how the model operates overall (Dieber & Kirrane, 2020). This local focus can be restrictive when a broader understanding of the model is needed (Saini & Prasad, 2022). Another limitation is its instability. Since LIME generates explanations based on random perturbations of input data, the explanations can vary from run to run, potentially leading to inconsistent insights (Dieber & Kirrane, 2020). Lastly, LIME can be computationally expensive to run, particularly for large datasets or complex models, as generating multiple perturbations and fitting local models for each prediction can be resource-intensive.

MPML leverages the same fundamental principle as LIME by locally approximating the model's behaviour around a specific prediction. However, MPML offers distinct advantages by allowing the interpretable model to be designed from scratch with interpretability built in, rather than relying on post-hoc approximation methods like LIME. While LIME constructs a simple, interpretable surrogate model for individual instances, MPML provides both local and global interpretability. Specifically, MPML delivers global insights into the ensemble's overall behaviour, which LIME does not. Moreover, MPML offers explanations in the form of impact scores at each layer of the ensemble, clearly outlining which groups of features had the greatest influence on the decision and, if necessary, identifying the specific features that contributed most to that outcome. This layered, structured interpretation offers more detailed and consistent insights compared to LIME's often variable, instance-level explanations. Nonetheless, both MPML and LIME share a common limitation, the computational overhead required to generate these explanations, which can be resource-intensive for complex models or large datasets.

2. SHAP (SHapley Additive exPlanations)

SHAP is a method in Explainable AI (XAI) that assigns importance scores to input features by using concepts from cooperative game theory, specifically the Shapley values (Li et al., 2024). These values represent the marginal contribution of each feature to a model's prediction by considering all possible combinations of feature subsets (Li et al., 2024; Merrick & Taly, 2020). SHAP provides a unified approach to interpreting predictions, which makes it model-agnostic and applicable to a wide range of machine learning algorithms (Aditya & Pal, 2022; Panda & Mahanta, 2023; Rodríguez-Pérez & Bajorath, 2020).

SHAP supports both global and local interpretability, allowing it to offer insights into how features generally affect the model as a whole, as well as explain individual predictions for specific instances (Aditya & Pal, 2022).

One of its major drawbacks is its computational complexity. Calculating Shapley values involves evaluating every possible feature combination, which can be computationally expensive, particularly for large datasets or complex models. While approximation methods exist to reduce the burden, they often come at the expense of precision. Another limitation is SHAP's assumption of feature independence. In real-world datasets, features often interact with each other, and SHAP may not fully capture these interactions, leading to potential inaccuracies in feature attribution. Lastly, due to the complexity of calculating feature contributions, SHAP can be resource-intensive, especially when applied to high-dimensional models or large datasets, requiring significant computational power and time.

MPML, like SHAP, assigns importance scores to input features that represent their contribution to a model's prediction by considering all possible combinations of feature subsets. This shared foundation allows both methods to capture complex feature interactions and provide detailed insights into model behaviour. Furthermore, both SHAP and MPML support global and local interpretability. The major difference however, is that SHAP can be applied post-hoc to any machine learning model, offering broad applicability, while MPML incorporates interpretability directly into the model's design.

Both approaches can be computationally intensive, especially when dealing with large feature spaces or complex models, due to the combinatorial nature of evaluating feature contributions. MPML's slight advantage, however, lies in its integration of interpretability within the model architecture itself, reducing reliance on external approximations and offering layer-wise impact scores that highlight not only individual feature contributions but also how groups of features influence decisions at different stages of an ensemble model. This structured approach can lead to more consistent and transparent explanations compared to SHAP's purely post-hoc analysis.

3. Bootstrap Aggregating

Bagging, short for Bootstrap Aggregating, is a well-established ensemble learning technique designed to improve model stability and predictive accuracy by reducing variance through model averaging. Introduced by Breiman (1996), bagging works by generating multiple bootstrap samples (random subsets of the original training data obtained with replacement) and training a separate model on each subset (Breiman, 1996). The predictions from these models are then combined, typically through majority voting for classification tasks or averaging for regression, to produce the final ensemble output. This approach enhances generalization performance by mitigating overfitting, especially for high-variance models like decision trees.

MPML, while sharing the ensemble philosophy of bagging, differs fundamentally in how diversity among ensemble components is introduced. Instead of creating different models by training on varying subsets of the data, MPML separates the input features into distinct groups, with each group being used to train a single model. This group-based feature partitioning emphasizes interpretability rather than relying on randomness in data sampling as in bagging. While bagging enhances predictive performance primarily through variance reduction and model averaging, MPML distinguishes itself by embedding interpretability directly within the ensemble architecture. This design not only improves the transparency of the model but also contributes to enhanced predictive accuracy.

4. Boosting

Boosting is a widely used ensemble learning technique designed to convert a collection of weak learners, models that perform only marginally better than random guessing, into a single strong learner capable of achieving high predictive accuracy (Mienye & Sun, 2022). The core principle behind boosting is the sequential training of models, where each subsequent model focuses on correcting the errors made by its predecessors. Popular boosting algorithms, such as AdaBoost and Gradient Boosting, assign higher weights to misclassified instances during the training process, ensuring that difficult examples receive increased attention in subsequent iterations (Bühlmann, 2012). Through this iterative error-correction mechanism, boosting reduces bias and improves overall model performance, making it highly effective for both classification and regression tasks. Despite its success, boosting tends to increase model complexity, which can reduce interpretability, particularly in applications involving high-dimensional data or intricate feature interactions.

MPML draws inspiration from boosting by leveraging the concept of multiple learners to improve predictive performance, but fundamentally differs in how these learners are constructed. Rather than building weak learners sequentially to iteratively correct the errors of previous models, as is characteristic of boosting (Mienye & Sun, 2022), MPML creates each learner by partitioning the input features into distinct groups. Each layer of the ensemble is dedicated to learning from a specific feature group, allowing the model to capture diverse patterns while maintaining a transparent structure. This design enables MPML to retain the performance benefits associated with ensemble methods while providing both global and local interpretability by clearly outlining the contribution of different feature groups to the final prediction. Thus, while MPML builds on the ensemble principles underlying boosting, it introduces a parallel, group-based learning framework that emphasizes interpretability without compromising accuracy.

5. The Need for Explainable AI in Healthcare

In healthcare, the demand for explainable AI (XAI) arises from several key reasons: ensuring regulatory compliance, addressing ethical concerns, and enhancing clinical outcomes. Understanding the inner workings of AI systems is essential for fostering trust and ensuring that these systems are managed and integrated into healthcare practice effectively. XAI provides the necessary transparency for clinicians, patients, and regulatory bodies to comprehend, trust, and oversee the decisions made by AI models (Amann et al., 2020).

XAI in healthcare can also facilitate meaningful interdisciplinary discourse among computer scientists, biomedical researchers, and clinicians, providing a shared framework for understanding complex model outputs and enabling collaborative decision-making to improve patient care outcomes. Omitting explainability in clinical decision support systems poses a threat to core ethical values in medicine and may have detrimental consequences for individual and public health. According to Adadi and Berrada (Adadi & Berrada, 2018), the need for XAI in healthcare can be linked to four primary reasons: justification, control, improvement and discovery.

Justification: Healthcare providers must justify AI-driven decisions to patients, especially in high-stakes scenarios such as diagnosis and treatment planning. XAI helps explain why a particular recommendation or diagnosis was made, allowing clinicians to provide evidence-based explanations to their patients and medical teams.

Control: In healthcare, controlling the outcomes of AI systems is vital to prevent harm and ensure patient safety. XAI empowers healthcare professionals by making the decision-making process of AI systems transparent, enabling them to identify and correct potential errors or biases in real time.

Improvement: Continuous improvement of AI systems is necessary to adapt to the evolving medical landscape. By making AI models explainable, healthcare professionals can better understand where the model may be lacking, allowing for iterative improvements that enhance accuracy and reliability over time.

Discovery: In healthcare, XAI can also serve as a tool for discovery. By revealing the underlying patterns and logic that AI systems use to make predictions, clinicians and researchers can gain new insights into medical data, potentially leading to novel scientific discoveries and innovations in patient care.

MPML is specifically designed to address the core requirements of explainable AI (XAI) in healthcare, making it well-suited for high-stakes, safety-critical environments. First, **Justification** is supported through MPML's ability to demonstrate the source of its decisions at multiple levels of abstraction. By separating features into distinct groups, or "perspectives," and providing impact scores at both the group and individual feature level, MPML offers clinicians transparent, structured explanations that clarify which factors contributed to a diagnosis or recommendation. Second, MPML enhances **Control** by enabling individual models created for each feature group to be independently altered, improved, or updated without requiring retraining of the entire ensemble. Third, MPML facilitates **Improvement** by allowing perspectives to be added, removed, or modified to enhance ensemble performance, all without retraining or reconstructing every base model. This flexibility supports continuous adaptation to new clinical data and evolving standards of care. Finally, MPML promotes **Discovery** by generating interpretable insights into how different groups of features—and specific variables within those groups—impact model predictions.

2. Multi-Perspective Machine Learning

Multi-Perspective Machine Learning (MPML) is an approach that integrates multiple perspectives of data to improve the accuracy and interpretability of Machine Learning models. In MPML, different subsets of features, often representing distinct aspects of the data, are modelled separately and then combined to provide a holistic prediction (Miller & Busby-Earle, 2017). This methodology not only enhances the robustness of the model by leveraging diverse data representations but also supports interpretability by allowing each perspective to be analysed independently. By focusing on the unique contributions of each perspective, MPML enables more nuanced insights into the model's decision-making process.

The Multi-Perspective Machine Learning (MPML) approach is devised to tackle a specific category of learning challenges, which exhibit the following characteristics: The ability to decompose the problem into distinct, independent components, facilitating a modular approach to problem-solving. The requirement for solutions to produce intelligible and transparent results, ensuring that outcomes are accessible and interpretable by stakeholders.

This ensemble methodology is particularly suited for addressing complex medical challenges such as heart disease, which has multiple independent facets that require distinct consideration. MPML is engineered to construct models that capture and represent the diverse aspects of the problem space.

2.1. Perspectives

The core component of this method is the perspective, a structured grouping of features that reflects a particular aspect or interpretation of the learning problem. To apply the approach effectively, each perspective must be clearly and thoughtfully defined. Perspectives are constructed using a variety of grouping strategies, including mutual information, model-based importance, correlation patterns, dimensionality reduction with clustering, and domain expert knowledge. These strategies enable the identification of coherent feature subsets that capture different dimensions of the data. The organization of features into perspectives allows the model to leverage distinct learning strategies, each tailored to a specific subset of information. This structure not only enhances interpretability but can also improve predictive performance (Zhao et al., 2017).

Importantly, the effectiveness of a given perspective depends on the nature of the dataset and the problem domain. For example, in the context of heart disease detection, one perspective may focus on clinical risk factors such as age, blood pressure, and cholesterol levels, while another may group imaging-based features derived from echocardiograms or cardiac MRIs (Johnson et al., 2018). By aligning feature groupings with distinct analytical approaches, this method supports both a modular model design and contextually grounded interpretation.

The MPML approach can be formally defined as follows:

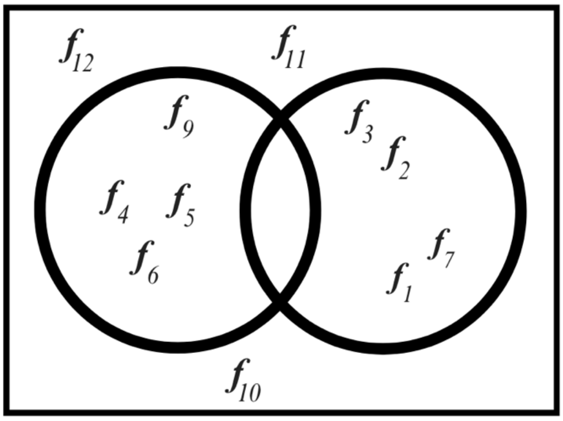

Figure 1: Learning Problem *T*

Let *T* represent a specific learning problem. Each element f_x in Figure 1 is a feature of the learning problem *T*.

$$T = \{f_1, f_2, f_3 \ldots f_n\}$$

Let *L* be the set of all possible learning strategies, l_x, that can be applied to solving problem *T* (Figure 2).

$$L = \{l_1, l_2, l_3 \ldots l_n\}$$

$$l_1 = \{f_1, f_2 \ldots f_x\}$$

$$\therefore\ l_x \subset T$$

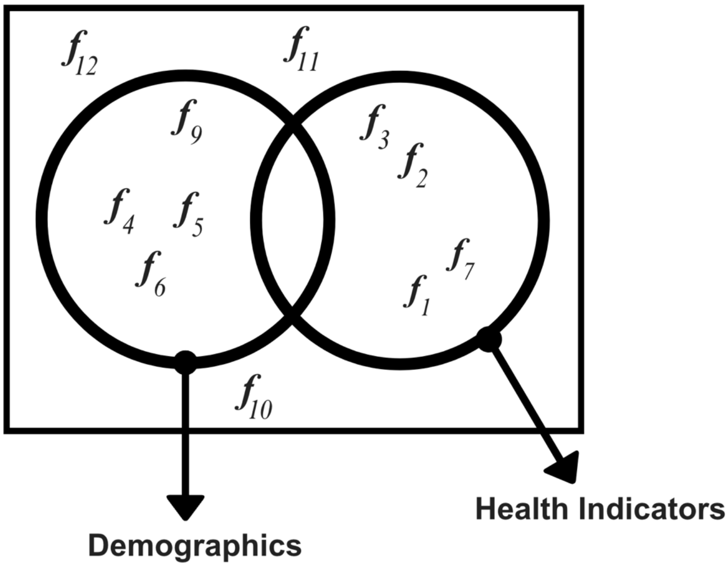

Figure 2: Learning Strategy *l*

Let P represent the set of perspectives of problem T. Within each learning strategy, there may be one or more subsets of features that describe specific aspects of the problem; these subsets we call perspectives.

$$P = \{p_1, p_2, p_3 \dots p_n\}$$

$$where \;\; p_x \subset T$$

$$and \;\; p_1 = \{f_1, f_2, \dots, f_3\}$$

These perspectives distinguish each classifier in the ensemble. The features from each perspective are used to create individual classifiers. Since each perspective comprises related features, the resulting classifiers are diverse. To achieve accuracy, each classifier is trained on the entire training set. The outcomes from each classifier are then combined to produce the final result. Every perspective belongs to a learning strategy. While a single learning strategy can include multiple perspectives, each perspective is associated with only one learning strategy (Figure 3).

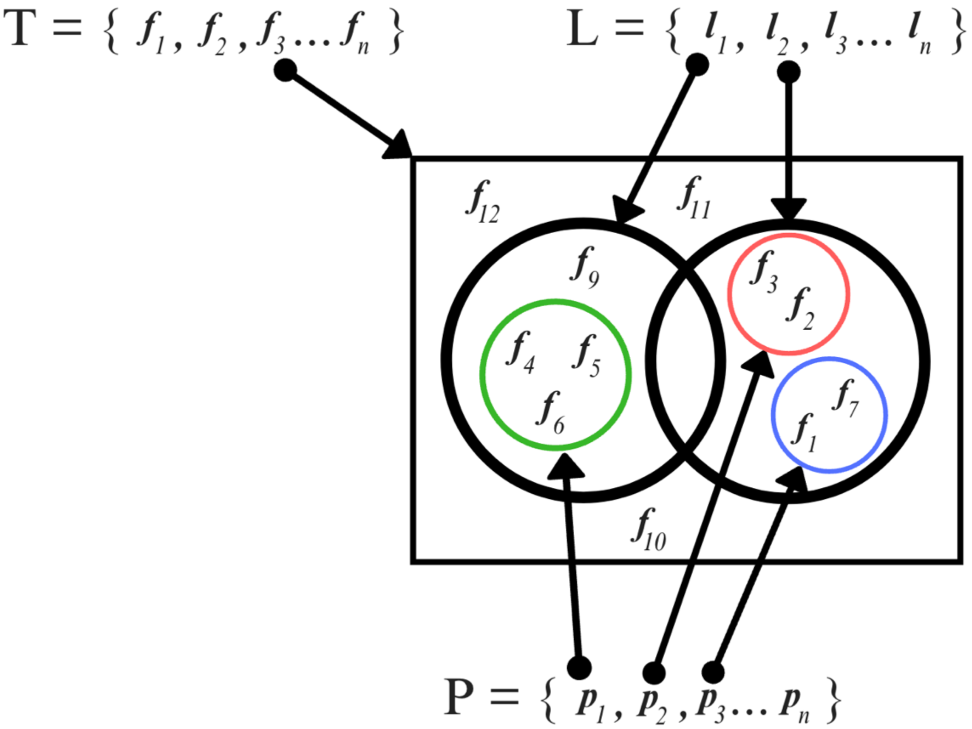
T = { f_1, f_2, f_3… f_n }
L = { l_1, l_2, l_3… l_n }
f_{12}
f_{11}
f_9
f_4
f_5
f_6
f_3
f_2
f_7
f_1
f_{10}
P = { p_1, p_2, p_3… p_n }

Figure 3: An Example MPML Breakdown

Example: Selecting Perspectives for Heart Disease Prediction

Let T be the learning problem, Heart Disease Prediction.

$$T = \{f_1, f_2, f_3 \cdots f_n\}$$

where each f_x is a feature used in Heart Disease Prediction:

$$\begin{aligned}
f_1 &= \mathit{Serum\ Cholestoral} \\
f_2 &= \mathit{Exercise\ induced\ angina} \\
f_3 &= \mathit{Resting\ electrocardiographic\ results} \\
f_4 &= \mathit{Age} \\
f_5 &= \mathit{Sex} \\
f_6 &= \mathit{Chest\ Pain\ Type} \\
f_7 &= \mathit{Resting\ Blood\ Pressure}
\end{aligned}$$

Each perspective p_1 (where $p_1 \in P$) is a subset of features from T that describes a portion of the problem task T. In this example we have,

$$\begin{aligned}
p_1 &= \{f_4, f_5, f_6\} - \mathit{Indicators} \\
p_2 &= \{f_2, f_3\} - \mathit{Diagnostic\ Features} \\
p_3 &= \{f_1, f_7\} - \mathit{Risk\ Factors}
\end{aligned}$$

Thus, the learning strategy l_1 and l_2 may be defined as:

$$\begin{aligned}
l_1 &= \{p_1\} \\
l_2 &= \{p_2, p_3\}
\end{aligned}$$

For each perspective p_x, a machine learning algorithm is applied to create a model with the features it contains. Each perspective thus forms a classifier (see Figure 4). These classifiers are then used to create an ensemble.

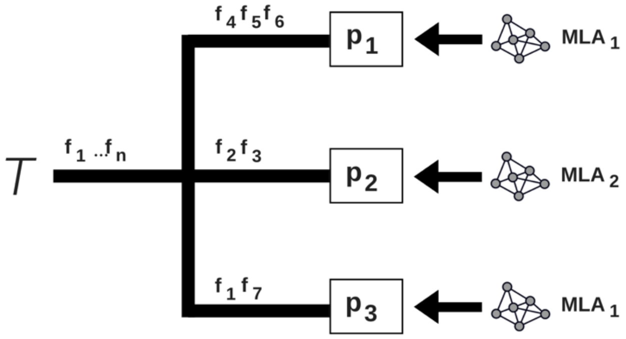

Figure 4: Applying Machine learning algorithm (MLA) to perspectives

An ensemble composed of classifiers derived from well-defined perspectives exhibits two essential properties for effectiveness: accuracy and diversity. The classifiers within an ensemble must not only be accurate, making reliable predictions individually, but also diverse, using different factors or features to predict outcomes. This diversity ensures that when their efforts are combined, the ensemble can make precise predictions and generalize well across different scenarios (Panhalkar & Doye, 2022). The model is finalized by employing an appropriate aggregation technique to combine the outputs of individual classifiers, yielding the final prediction. In the case of MPML, the default combination strategy utilized in this study is blending.

2.2. Feature Grouping Methods

Several methods for constructing perspectives were explored, incorporating both data-driven and expert-informed strategies to ensure a balance between empirical structure and clinical interpretability. The following grouping approaches were selected for this study due to their complementary strengths. Mutual Information (MI)–based grouping was used to identify features with strong dependency relationships to the target variable, enabling perspectives that capture direct predictive relevance. Model importance–based grouping leverages feature-importance scores from tree-based models to cluster variables that contribute similarly to prediction, offering a pragmatic, model-aware structure. Correlation-based grouping supports the identification of features that behave similarly across samples, reducing redundancy and creating perspectives grounded in statistical coherence. Dimensionality-reduction and clustering methods were included to uncover latent structure and natural groupings within the data, allowing the framework to detect relationships that may not be obvious through univariate measures. Finally, domain expert–defined grouping was incorporated to ensure that perspectives reflect clinically meaningful constructs, aligning the model with established cardiovascular knowledge. Together, these methods were chosen to provide a robust and diverse set of perspectives that balance interpretability, data-driven insight, and methodological rigour.

Mutual Information (MI)-Based Grouping: This method groups numeric features based on their mutual information (MI), which reflects how much information one feature shares with another. It selects only numeric features and discretizes them into bins using a specified strategy (e.g., uniform, quantile, or k-means), with the number of bins determined by Sturges' Rule. The method then computes pairwise mutual information scores between all features to assess their informational similarity. These scores are normalized and converted into a distance matrix, which is used as input for agglomerative hierarchical clustering. The features are then clustered into a user-defined number of groups.

Model Importance-Based Grouping: This method groups numeric features based on their importance in predicting a target variable, using a tree-based model (in this case, a Random Forest). After training, it extracts feature importance scores, which indicate how much each feature contributes to the model's predictive accuracy. These scores are then standardized and clustered using k-means into a specified number of groups.

Correlation-Based Grouping: This method groups numeric features based on the similarity of their correlation patterns. It begins by computing a correlation matrix using the Pearson method. The absolute values of the correlations are taken and subtracted from 1, so that highly correlated features have smaller distances. This distance matrix is then converted into a condensed form suitable for hierarchical clustering. Using average linkage, the features are hierarchically clustered, and a flat clustering is produced based on the desired number of groups.

Dimensionality Reduction and Clustering: This approach groups numeric features by first projecting them into a lower-dimensional space using a dimensionality reduction technique, and then applying clustering to identify groups of similar features. The data is transposed so that each feature becomes a sample, allowing the algorithm to analyse relationships between features rather than between data points. These transposed features are standardized and then projected into a lower-dimensional space using a user-specified method: PCA (Principal Component Analysis). This dimensionality reduction step captures the main patterns in feature variation. The projected data is then clustered using k-means, and each original feature is mapped to a cluster, resulting in interpretable, similarity-based feature groups.

Domain Expert-Defined Grouping: This method organizes features into meaningful subsets based on the knowledge and judgment of subject matter experts. Unlike statistically derived perspectives, which rely on algorithmic criteria such as mutual information, correlation patterns, or variance structure, the expert-defined perspective introduces an intentionally subjective, human-guided dimension to the framework. Its purpose is not only to reflect clinical, conceptual, or operational relevance but also to serve as a contrast against more formal, data-driven selection methods. By incorporating expert reasoning directly into the model design, this perspective functions as a human-centric control mechanism, allowing us to examine how domain insight affects performance relative to purely statistical approaches. This enhances interpretability and provides a valuable benchmark for understanding when, and to what extent, expert intuition complements or diverges from algorithmic feature selection.

2.3. Generating Interpretations with MPML

The structure of the method is illustrated in Figure 5, which shows a typical MPML setup with interpretation possible at each level. For any given instance or patient, the system can provide an explanation for the prediction by identifying the perspective with the greatest impact score and by reporting the features that have the highest individual impact scores within each perspective.

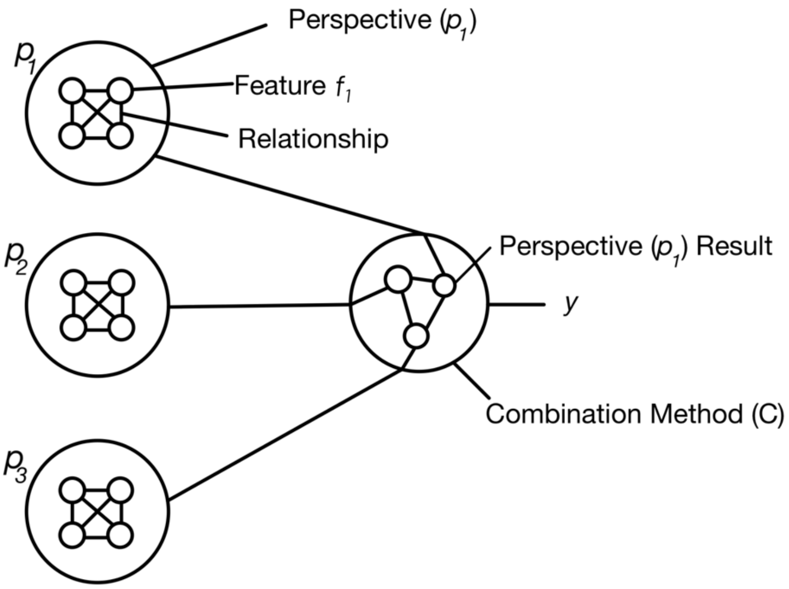

Figure 5: Model Overview

Each perspective (p_x) focuses on a single aspect of the learning problem (T). Understanding how each perspective affects the prediction y provides a specific interpretation for that particular instance. For example (see Figure 6), if perspective one (p_1) is the most influential in predicting heart disease for a patient, and this perspective (p_1) is composed of diagnostic features, then this insight offers valuable information about the patient's cardiac function or patterns that contribute to the diagnosis.

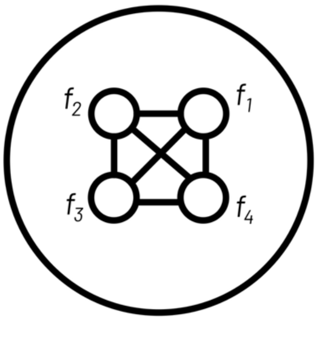

Figure 6: A single Perspective p_1

By delving deeper than the perspective level, we can identify the most influential features within the most impactful perspective. This deeper analysis provides insight into which specific diagnostic features are most effective for predicting heart disease. Understanding how these individual features relate to each other is crucial for comprehending the underlying behaviour of the condition. The formal steps to obtain the impact score for each feature within a perspective are defined as follows:

P is the set of all perspectives for a given learning Task T.

$$P = \{p_1, p_2, p_3 \ldots p_n\}$$

Each perspective p_x contains a subset of features f_x from the learning task T.

$$p_x = \left\{f_1, f_2, f_3 \cdots f_n\right\}$$

The model generated by applying a learning algorithm S to any perspective p_x is represented as,

$$S_x(p_x)$$

The set of all models S produced from each perspective in P is denoted by Q

$$Q = (S_1\left(p_1\right),\ S_2\left(p_2\right),\ S_3\left(p_3\right) \ldots S_n(p_n))$$

These models are then combined using a combination method C and the final result (the prediction) is represented by y

$$y = \ C(S_1\left(p_1\right),\ S_2\left(p_2\right),\ S_3\left(p_3\right) \ldots S_n(p_n))$$

$$y = \ C(Q)$$

We then aim to explain y using a method similar to the EXPLAIN technique (Robnik-Sikonja & Kononenko, 2008) by identifying which perspective, when removed, causes the greatest change in y.

To calculate the change in y, we examine the model's confidence in its prediction of y. For example, in heart disease prediction, the result y could be either 1 (indicating heart disease) or 0 (indicating no heart disease). We record the model's confidence for each class. If the model (with all perspectives) predicts a 1 with 90% confidence and a 0 with 10%, we note the confidence in the correct class, which is 90%. After removing a perspective ($S_x(p_x)$), the new result is stored as $\hat{y}$. If the model now has 70% confidence that the result is 1 and 30% confidence that it is 0, the change in y, stored as d, would be $90 - 70 = 20$.

Perspective $(S_x(p_x))$ has the greatest impact on y if the resulting $\hat{y}$, computed without $(S_x(p_x))$, shows the largest difference from y across all perspectives $p_x \in P$.

$$\hat{y} = C(Q - \{S_x(p_x)\})$$
$$d = y - \hat{y}$$

This process is repeated from the output y until the most influential feature within the strongest perspective is identified. The value of d also indicates the direction the model moves when $S_x(p_x)$ is removed. If removing $S_x(p_x)$ brings $\hat{y}$ closer to the correct prediction, then $S_x(p_x)$ negatively impacts the result y for that specific case. If the opposite happens, and removing $S_x(p_x)$takes the prediction further from the correct result, then $S_x(p_x)$ has a positive impact on y. Both positive and negative impacts are helpful in providing clinicians with a clearer understanding of the model's behavior and determining whether it can be trusted for use. The impact score d quantifies the contribution of individual features to the model's predictions, providing deeper insights into the factors driving the model's decision-making process. This metric is instrumental for both local and global interpretability, enabling a more comprehensive understanding of the model's behavior.

3. Methodology

3.1. Datasets Preparations

The methodological process for this study begins with the preparation and loading of two cardiovascular datasets that serve as inputs to the model evaluation pipeline. The primary dataset used for the MPML experiments is a curated, comprehensive heart disease dataset constructed by merging several well-known clinical datasets: the Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog Heart Disease datasets. These five datasets are commonly referenced in cardiovascular risk–prediction literature and collectively provide a mixture of demographic, clinical, and diagnostic variables relevant to heart disease classification. The merged dataset contains 1,190 instances compiled across 11 shared features, making it the largest publicly available structured heart disease dataset constructed from these sources. The motivation for using this dataset lies in its breadth and its historical relevance for benchmarking machine learning approaches in cardiovascular prediction tasks. However, the integration of multiple datasets inevitably introduces heterogeneity arising from differences in population distributions, diagnostic practices, hospital systems, measurement protocols, and label conventions. While these factors can influence absolute model performance, it is important to clarify that the present study does not aim to evaluate the dataset itself, nor to make claims about the clinical validity of predictive outcomes. Rather, the dataset's role is to serve as a standardized input for systematically evaluating the proposed Multi-Perspective Machine Learning (MPML) framework against conventional ensemble-learning baselines. Therefore, although dataset heterogeneity exists, its effects are controlled by applying identical preprocessing, splits, and evaluation procedures across all modeling approaches. This ensures that the comparison reflects differences in modeling frameworks rather than differences in data composition.

In addition to the merged heart disease dataset, the study also employs a secondary cardiovascular dataset consisting of 70,000 patient records (34,979 presenting with cardiovascular disease and 35,021 not presenting with cardiovascular disease) with 11 features collected during routine medical examinations. This dataset includes objective measurements such as age, height, weight, and blood pressure; examination-derived indicators such as cholesterol and glucose levels; and subjective lifestyle factors such as smoking, alcohol consumption, and physical activity. Each record includes a binary label indicating the presence or absence of cardiovascular disease. The dataset was designed to capture a broader clinical and behavioral profile of heart health, and its structure makes it suitable for evaluating perspective-level modeling, particularly in the probability-calibration experiments where a single perspective is isolated to study calibrated outputs. As with the first dataset, this secondary dataset is not evaluated as the subject of inquiry. Its purpose is exclusively methodological: it provides an alternative feature distribution and clinical framing through which to test MPML's interpretability mechanisms and error-analysis procedures. Across the entire study, datasets are treated as controlled experimental inputs rather than as objects of scientific evaluation, ensuring alignment with the paper's central goal of demonstrating and analyzing the MPML framework.

3.2. Model Preparation

The methodological process in this study begins with loading the heart disease dataset into a pandas Python DataFrame, separating it into features and the class label, and applying a consistent 70/30 train–test split. This split is maintained across all experiments for comparability, while 10-fold cross-validation is introduced for more robust performance estimation. The numerical nature of the dataset allows direct use without additional encoding steps, and random seeds are fixed to ensure reproducibility. All models—MPML and the baseline ensembles—are trained under standardized experimental conditions to isolate the effect of the algorithmic design rather than preprocessing differences.

The Multi-Perspective Machine Learning (MPML) framework (Miller & Busby-Earle, 2017) is then applied as the primary experimental approach. MPML begins by organizing features into predefined groups, or perspectives, representing distinct conceptual dimensions within the dataset such as physiological measures, demographic attributes, or diagnostic indicators. These perspectives remain fixed throughout training and are intentionally kept manually defined rather than algorithmically generated to emphasize interpretability and domain-awareness. The MPML ensemble is initialized using a decision-tree base estimator with blending as the ensemble strategy and a meta-integration rule that fuses predictions across perspectives. The rationale for selecting a decision tree as the base learner lies in its interpretability, low computational cost, and ability to model non-linear relationships. Trees also complement MPML structurally, because each perspective is low-dimensional, making complex models unnecessary and potentially counterproductive. No hyperparameter tuning is performed on the base estimator because the intention behind MPML is to evaluate the power of feature-perspective decomposition rather than parameter optimization. Thus, the baseline tree configuration is intentionally simple, ensuring that any performance gain arises from the MPML architecture rather than deep algorithmic tuning.

Perspective generation follows, during which MPML extracts the appropriate feature subsets for each group and constructs perspective-specific datasets. This step operationalizes MPML's core concept by allowing multiple specialized models to learn from coherent feature subsets rather than the full feature space. The approach prioritizes interpretability over aggressive hyperparameter tuning, reflecting MPML's design philosophy: improving performance not by increasing model complexity, but by structuring feature information more effectively. A custom 10-fold cross-validation procedure is then applied, where for each fold the ensemble trains one model per perspective and blends predictions to produce a final decision. Performance metrics—including accuracy, precision, recall, and F1 score—are recorded for every fold and averaged to obtain the final MPML evaluation. Because the purpose of MPML in this study is architectural evaluation, no advanced tuning such as depth restrictions, pruning, or parameter searches is performed; the selected settings maintain transparency and ensure that the comparison emphasizes the MPML method rather than model-specific optimization.

To contextualize MPML's performance, three classical ensemble models—Random Forest, Bagging, and Gradient Boosting—are implemented as baseline comparators. These models are configured using modest and interpretable hyperparameter values. For Random Forest, the number of trees is set to 23, balancing computational efficiency with stability; this choice is deliberately medium-sized, avoiding both overly small ensembles and unnecessarily large forests that complicate interpretability. Bagging uses 16 decision-tree estimators, reflecting the principle that bagging primarily stabilizes variance and therefore does not require excessive model counts for datasets of this scale. Gradient Boosting employs 23 boosting stages, a conservative configuration meant to evaluate the model in a standard form rather than a highly optimized state. Across all three ensemble methods, hyperparameters are intentionally kept close to typical defaults to ensure that the models serve as baseline comparisons rather than optimized competitors. No grid search or parameter tuning is performed because the purpose is not to identify the best possible classical ensemble, but rather to benchmark MPML against commonly used, reasonably configured models whose performance reflects their general characteristics rather than hyperparameter engineering.

Each baseline ensemble is evaluated using stratified 10-fold cross-validation, ensuring consistent class distributions across folds. In each fold, the model trains on nine subsets and predicts the tenth, producing fold-level performance metrics identical to those computed for

MPML. The use of cross-validation rather than a single train–test split ensures reliability and reduces sensitivity to sample variation. By keeping the tuning minimal and transparent, the study avoids overfitting the baselines and allows for a fair conceptual comparison: MPML's structural advantages versus the traditional ensembles' standard learning mechanisms.

3.3. Interpretation Analysis

A final experimental component addresses interpretability and probability estimation through calibration analysis applied to a decision-tree model trained on a single MPML perspective. Three calibration settings are evaluated: the raw uncalibrated tree, Platt scaling via a sigmoid transformation, and isotonic regression. These calibration methods are chosen because they represent the two most widely used probability-adjustment techniques in machine learning—one parametric and one nonparametric. No tuning beyond default parameters is applied because the purpose is to illustrate how probability distributions change under different calibration rules, not to maximize predictive accuracy. A specific test instance is examined to compare probability outputs across calibration methods, demonstrating how calibrated models adjust confidence even when the predicted class remains consistent. The study also identifies and inspects misclassified instances within the test set, extracting feature profiles and predicted labels for up to five incorrectly classified cases. This qualitative inspection supports the interpretability goals of MPML and provides additional insight into model behavior beyond aggregate statistics.

The methodology combines conservative, transparent baseline configurations with a structured MPML modeling approach to ensure that any observed performance differences arise from the intrinsic design of the methods rather than from aggressive hyperparameter optimization. This methodological choice prioritizes clarity, fairness, and interpretability, aligning with the study's objective to assess MPML as a conceptual modeling framework rather than a parameter-tuned optimization exercise.

4. Datasets

In this section, we provide a detailed overview of datasets used to compare the proposed approach (MPML) to other ensemble techniques. Two datasets were utilized for this study: a combined heart disease dataset and a cardiovascular disease dataset. The inclusion of the cardiovascular disease dataset was specifically intended to assess the scalability and robustness of the methods on a significantly larger dataset.

The heart disease dataset contains just over 1,000 instances, while the cardiovascular disease dataset comprises more than 70,000 instances, providing a comprehensive evaluation of each method's performance across datasets of varying sizes. This setup ensures that the comparison between MPML and other ensemble methods reflects not only general predictive capability but also adaptability to different data scales.

4.1. Heart Disease Dataset

Table 1: Heart Disease Dataset Feature Descriptions

No.	Feature	Code	Type	Description
1	Age	age	Numeric	Age in years
2	Sex	sex	Binary	1 = male, 0 = female
3	Chest Pain Type	chest pain type	Nominal	1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic
4	Resting Blood Pressure	resting bp	Numeric	Resting blood pressure (in mm Hg)
5	Serum Cholesterol	cholesterol	Numeric	Cholesterol level (in mg/dl)
6	Fasting Blood Sugar	fasting blood sugar	Binary	1 = true (> 120 mg/dl), 0 = false
7	Resting ECG Results	resting ecg	Nominal	0 = normal, 1 = ST-T abnormality, 2 = left ventricular hypertrophy
8	Max Heart Rate Achieved	max heart rate	Numeric	Maximum recorded heart rate
9	Exercise-Induced Angina	exercise angina	Binary	1 = yes, 0 = no
10	ST Depression	oldpeak	Numeric	Depression relative to rest
11	ST Slope	ST slope	Nominal	1 = upsloping, 2 = flat, 3 = downsloping
12	Heart Disease (Target)	class	Binary	1 = heart disease, 0 = no heart disease

For this study, we utilized a comprehensive heart disease dataset created by combining five widely used but previously independent datasets (Alizadehsani et al., 2019; Manu Siddhartha, 2025). This curated dataset represents one of the most extensive publicly available resources for coronary artery disease (CAD) prediction using machine learning techniques (see Table 1). The integration of these datasets allows for a more diverse and representative collection of instances, enhancing the robustness and generalizability of the experimental evaluation.

The merged dataset consists of 1,190 instances and includes 11 clinically relevant features that are consistent across all source datasets. These features were selected to ensure compatibility and meaningful analysis across the combined dataset. The five datasets used for this integration are: Cleveland, Hungarian, Switzerland, Long Beach VA, and the Statlog (Heart) Data Set. This unified dataset has been widely recognized in the literature for its suitability in developing and benchmarking machine learning models for CAD prediction.

4.2. Cardiovascular Disease dataset

In this study, we also utilized the Cardiovascular Disease dataset, obtained from Kaggle employed by several peer-reviewed articles (Ali, 2025; Saridena et al., 2023). The dataset contains data from 70,000 patient records collected during routine medical examinations (see Table 2). This large-scale dataset is designed to support the development of predictive models for cardiovascular disease detection. It consists of 11 input features along with a binary target variable that indicates the presence or absence of cardiovascular disease in each patient. All feature values were recorded at the time of examination, providing a consistent and reliable dataset suitable for machine learning applications.

Table 2: Cardiovascular Disease Dataset Feature Descriptions

No.	Feature	Code	Category	Description
1	Age	age	Objective	Age of the patient in days
2	Height	height	Objective	Patient's height in centimetres
3	Weight	weight	Objective	Patient's weight in kilograms
4	Gender	gender	Objective	Gender (coded as categorical values)
5	Systolic Blood Pressure	ap_hi	Examination	Systolic arterial pressure
6	Diastolic Blood Pressure	ap_lo	Examination	Diastolic arterial pressure
7	Cholesterol Level	cholesterol	Examination	1 = normal, 2 = above normal, 3 = well above normal
8	Glucose Level	gluc	Examination	1 = normal, 2 = above normal, 3 = well above normal
9	Smoking Status	smoke	Subjective	1 = smokes, 0 = does not smoke
10	Alcohol Intake	alco	Subjective	1 = consumes alcohol, 0 = does not consume alcohol
11	Physical Activity	active	Subjective	1 = physically active, 0 = not physically active
12	Cardiovascular Disease (Target)	cardio	Target	1 = has cardiovascular disease, 0 = no disease

The features in the Cardiovascular Disease dataset are grouped into three main categories. The first category, Objective Features, includes direct factual information such as age, height, weight, and gender. The second category, Examination Features, encompasses clinical measurements collected during the examination, including blood pressure, cholesterol levels, and glucose levels. The final category, Subjective Features, captures self-reported behaviours and lifestyle factors such as smoking status, alcohol consumption, and levels of physical activity. This combination of objective, clinical, and behavioural data allows for a comprehensive analysis of factors contributing to cardiovascular disease risk.

5. Experiments and Results

In this section, we present a comprehensive overview of the experimental setup, including the evaluation procedures and methodologies employed to compare the MPML approach against established ensemble techniques. A performance comparison was conducted to assess MPML under various configurations in relation to other ensemble models. In addition to standard performance evaluation, a series of paired t-tests were performed across multiple metrics to rigorously assess the statistical significance of any observed differences between MPML and the baseline ensemble methods.

Furthermore, this section provides a detailed presentation of the experimental results, supported by tables and thorough explanations, to offer a clear interpretation of the outcomes and validate the effectiveness of the proposed approach.

5.1. Experiments

Performance Comparison:

In this evaluation, multiple classification methods were tested and compared using key performance metrics, including Accuracy, F1 Score, Precision, and Recall. The methods included traditional classifiers such as Naive Bayes, Decision Tree, and Support Vector Machine (SVM), as well as ensemble techniques like Bagging, Boosting, and Random Forest. Additionally, several configurations of the MPML (Multi-Perspective Machine Learning) approach were assessed. MPML leverages different feature selection techniques, including mutual information (mi), correlation analysis, Principal Component Analysis (PCA), model-based importance ranking, and expert-defined feature groups, either individually or in combination.

The performance of these MPML configurations was compared at different ensemble sizes—specifically using 4, 7, 16, and 23 base classifiers—to analyse how ensemble complexity impacts results. Similarly, the number of base classifiers for Bagging, Boosting, and Random Forest was adjusted to align with MPML's varying ensemble sizes to ensure fair, consistent comparisons. For the ensemble methods, Decision Trees were used as base estimators where applicable. The experiments were designed to comprehensively evaluate how individual and combined feature selection strategies within MPML compare to traditional machine learning models and established ensemble approaches across multiple performance dimensions.

Paired t-Test Comparison:

In this evaluation, a series of paired t-tests were conducted to statistically compare the performance of the MPML ensemble method against three well-known ensemble techniques: Boosting, Bagging, and Random Forest. The MPML approach integrates multiple feature selection strategies, including mutual information (mi), correlation analysis, Principal Component Analysis (PCA), model-based importance measures, and expert-defined feature groups, resulting in an ensemble of 23 base classifiers. For comparison, Boosting was implemented using a *GradientBoostingClassifier* with 100 estimators, Bagging utilized a *BaggingClassifier* with 100 *DecisionTreeClassifier* estimators, and *Random Forest* was configured with 1,000 decision trees.

The configurations for Boosting, Bagging, and Random Forest were not chosen arbitrarily; rather, the number of base classifiers for each method was determined after conducting multiple experimental runs on the same dataset to identify the most effective configuration in terms of predictive performance. These optimized settings were then used in the final comparison to ensure a fair and meaningful evaluation against MPML.

The tests were performed across four key performance metrics: Accuracy, Precision, Recall, and F1 Score. In each case, a paired t-test was applied to assess whether the observed differences between MPML and the other ensemble methods were statistically significant. The comparisons were based on results obtained through 10-fold cross-validation, ensuring that each model was evaluated on multiple training and testing splits of the dataset. This approach provides a robust, unbiased estimate of performance and strengthens the reliability of the statistical conclusions drawn from the t-tests regarding the relative effectiveness of MPML compared to the other ensemble methods.

5.2. Results with Heart Disease Dataset

Performance Comparison (with Heart Disease Dataset)

Table 3 provides a comprehensive performance comparison between traditional machine-learning classifiers, standard ensemble methods, and a variety of MPML (Multi-Perspective Machine Learning) configurations. The results highlight clear performance stratification across methods and demonstrate the substantial benefits of the MPML framework, particularly when multiple complementary perspectives are combined. Among the baseline models, Naive Bayes and Decision Trees perform reasonably well, achieving accuracies of 0.857 and 0.870 respectively, while SVM lags behind with an accuracy of 0.726 and the lowest F1 score in the table. These results reinforce the well-known limitations of SVM under certain feature distributions and class-balance conditions. Traditional ensemble methods substantially improve upon these baselines: Bagging (using 100 Decision Trees) achieves an accuracy of 0.931, Boosting reaches 0.882, and Random Forest (100 trees) delivers strong overall performance with an accuracy of 0.947 and a recall of 0.963, making it the strongest of the non-MPML models.

The MPML configurations introduce a different layer of analysis by strategically selecting and combining base classifiers based on expert knowledge, feature relationships, and model-driven metrics. Even the simpler MPML setups—such as those based on mutual information, correlation filtering, PCA, or model importance—perform competitively, with accuracies ranging from 0.856 to 0.894 using only four base classifiers. Notably, the MPML Expert Groups configuration, which incorporates seven carefully selected base classifiers, achieves an accuracy of 0.926 and balanced precision–recall performance. This places it in the same range as Bagging with 100 estimators, but with far fewer base models, highlighting MPML's efficiency through strategic selection rather than brute-force ensembling.

The most advanced MPML configurations clearly outperform all other models in the table. The combination of mutual information + correlation + PCA + model importance achieves an accuracy of 0.966 with only 16 base classifiers, surpassing even the 100-tree Random Forest.

Table 3: Performance Comparison on Heart Disease Dataset

Method	Accuracy	F1 Score	Precision	Recall	Base Classifiers
Naive Bayes	0.857	0.875	0.869	0.882	N/A
Decision Tree	0.870	0.882	0.915	0.851	N/A
SVM	0.726	0.745	0.790	0.704	N/A
Bagging (estimator=DecisionTree)	**0.931**	**0.935**	**0.930**	**0.941**	100
Boosting	**0.882**	**0.889**	**0.891**	**0.887**	100
Random Forest	**0.947**	**0.951**	**0.938**	**0.963**	100
MPML (Expert Groups)	**0.926**	**0.926**	**0.928**	**0.927**	7
MPML (model_importance)	**0.894**	**0.893**	**0.894**	**0.894**	4
MPML (PCA)	0.856	0.853	0.863	0.856	4
MPML (corelation)	0.885	0.883	0.885	0.885	4
MPML (mi)	0.879	0.877	0.878	0.879	4
MPML (mi + corelation + PCA + model_importance)	**0.966**	**0.966**	**0.967**	**0.966**	16
MPML (mi + corelation + PCA + model_importance + Expert Groups)	**0.955**	**0.954**	**0.954**	**0.955**	23
Bagging (estimator=DecisionTree)	0.912	0.916	0.920	0.913	7
Boosting	0.839	0.854	0.824	0.889	7
Random Forest	0.913	0.918	0.914	0.922	7
Bagging (estimator=DecisionTree)	0.921	0.925	0.930	0.921	16
Boosting	0.849	0.859	0.848	0.873	16
Random Forest	0.934	0.937	0.934	0.941	16
Bagging (estimator=DecisionTree)	0.934	0.938	0.932	0.944	23
Boosting	0.850	0.860	0.851	0.871	23
Random Forest	0.938	0.942	0.932	0.952	23

When expert knowledge is added to this composite configuration—resulting in the 23-classifier MPML ensemble—the model achieves extraordinarily high performance across all metrics (Accuracy = 0.955, F1 = 0.954, Precision = 0.954, Recall = 0.955). Although slightly lower than the 16-classifier configuration, this variant remains one of the strongest overall and demonstrates that incorporating expert-driven perspective selection maintains high model stability and generalization quality.

When comparing MPML configurations directly against traditional ensemble methods using matched numbers of base classifiers, the performance advantage becomes even more pronounced. With seven base classifiers, MPML Expert Groups (Accuracy = 0.926) outperforms Bagging (0.912), Boosting (0.839), and Random Forest (0.913). With sixteen classifiers, the disparity increases: the MPML composite model achieves an accuracy of 0.966, compared with Bagging at 0.921, Boosting at 0.849, and Random Forest at 0.934. At 23 base classifiers, MPML again leads, outperforming Bagging (0.934), Boosting (0.850), and Random Forest (0.938). These consistent gains highlight the strength of the MPML methodology, which combines multiple feature-selection perspectives to build ensembles that are not only more accurate but also more balanced across precision, recall, and F1 score.

The results in Table 3 reinforce the central value of MPML's multi-perspective design philosophy. By leveraging complementary feature signals—such as mutual information, correlation structure, principal components, and model-derived importance rankings—MPML produces ensembles that systematically outperform both traditional machine-learning models and conventional, single-strategy ensemble methods. The ability to achieve such high accuracy with relatively few base classifiers underscores MPML's efficiency and its potential to provide more interpretable, computationally tractable, and high-performing solutions in real-world classification tasks.

Paired t-Test Comparison (with Heart Disease Dataset)

The comparative evaluation of the MPML ensemble method against established ensemble techniques demonstrates the statistically significant superiority of MPML across multiple performance metrics. Using paired t-tests, the MPML configuration, which integrates mutual information (mi), correlation, principal component analysis (PCA), model importance, and expert grouping, consistently outperforms its counterparts in accuracy, precision, recall, and F1 score, with all p-values being effectively zero (or ≤ 0.0001), indicating high statistical significance. Despite employing only 23 base classifiers, MPML yielded t-statistics as high as 18.1215 (F1 score vs. Boosting) and 17.4815 (accuracy vs. Boosting), surpassing Boosting, Bagging, and Random Forest models that utilize substantially more base classifiers (100 to 1000).

This performance highlights the effectiveness of MPML's diverse and strategically selected ensemble design over traditional methods, which rely primarily on high estimator counts and do not incorporate the same depth of feature selection and expert-informed grouping. The consistent dominance across all metrics supports the robustness and generalizability of the MPML framework.

McNemar test comparisons (with Heart Disease Dataset)

The set of McNemar test comparisons in Table 4 provides a detailed picture of how different MPML ensemble configurations perform relative to Random Forest baselines. When comparing MPML Stacking using Gaussian Naive Bayes as the meta-model against a Random Forest with 23 estimators, the results show no meaningful performance difference between the two approaches. The off-diagonal counts—8 instances where the stacking model is correct while the Random Forest is wrong, versus 10 instances where the Random Forest is correct and the stacking model is wrong—are nearly symmetrical. This balance is confirmed by the very high p-value (0.8145), indicating that any observed differences are well within the range of random variation. In practical terms, the two models can be considered statistically equivalent for this dataset, meaning the choice between them should depend on secondary factors such as interpretability, computation time, or deployment simplicity rather than predictive superiority.

Table 4: Summary of McNemar Test Results for MPML Models vs. Random Forest Baselines

Comparison	MPML (Correct)	RF (Correct)	p-value	Significance	Better Performing Model
MPML Stacking (GaussianNB) 23 vs. Random Forest 23	8	10	0.8145	Not significant	None (models equivalent)
MPML Stacking (DT) 23 vs. Random Forest 23	12	25	0.047	Significant (p < 0.05)	Random Forest 23
MPML Blending (DT) 23 vs. Random Forest 23	20	8	0.036	Significant (p < 0.05)	MPML Blending (DT)
MPML Blending (DT) 23 vs. Random Forest 1000	16	3	0.0044	Highly significant (p < 0.01)	MPML Blending (DT)

In contrast, the comparison between MPML Stacking with a Decision Tree (DT) meta-model and the same Random Forest (23 estimators) reveals a statistically significant difference in performance. Here, the Random Forest model proves to be more accurate, with 25 cases where it correctly predicts while the stacking model does not, compared to only 12 cases in the opposite direction. With a p-value of approximately 0.047, this difference crosses the threshold for statistical significance and suggests that the Random Forest is the more reliable model among the two.

However, this advantage does not hold in the blending-based comparisons. When evaluating MPML Blending (DT) vs. Random Forest (23 estimators), the direction of superiority reverses. The blended model correctly classifies 20 cases that the Random Forest gets wrong, while the Random Forest outperforms the blended model in only 8 instances. The resulting p-value (≈ 0.036) shows that this difference is statistically significant, implying that the blended model provides a meaningful improvement over the Random Forest under these conditions.

This trend becomes even more pronounced when comparing MPML Blending (DT) against a much larger Random Forest with 1000 estimators. Despite the increased complexity and capacity of the larger Forest, the blended model still demonstrates significantly better performance, with 16 unique correct predictions compared to only 3 for the Random Forest. The highly significant p-value (≈ 0.0044) reinforces that the blended model offers a substantial and reliable performance advantage.

Overall, these results illustrate how different ensemble strategies—stacking vs. blending, GaussianNB vs. Decision Tree meta-models—can vary widely in effectiveness depending on the configuration. While some MPML variants match the performance of traditional models, others outperform Random Forests even when the latter are scaled to a much larger size. These tests demonstrate the value of using statistically rigorous pairwise comparison methods like McNemar's test, as they reveal not just differences in overall accuracy but meaningful differences in error patterns, enabling a more informed selection of models for deployment.

5.3. Results with Cardiovascular Disease dataset

Performance Comparison (with Cardiovascular Disease Dataset)

Table 4: Performance Comparison on Cardiovascular Disease Dataset

Method	Accuracy	F1 Score	Precision	Recall	Base Classifiers
Naive Bayes	0.595	0.444	0.713	0.323	N/A
Decision Tree	0.635	0.638	0.634	0.642	N/A
SVM	0.605	0.589	0.616	0.563	N/A
Bagging (estimator=DecisionTree)	**0.714**	**0.712**	**0.721**	**0.702**	100
Boosting	**0.738**	**0.73**	**0.754**	**0.708**	100
Random Forest	**0.717**	**0.714**	**0.725**	**0.705**	100
MPML (cardio_all_expert_grouping)	**0.772**	**0.771**	**0.775**	**0.772**	20
MPML (cardio_expert_and_stat_grouping)	**0.852**	**0.851**	**0.858**	**0.852**	36
Bagging (estimator=DecisionTree)	0.706	0.699	0.718	0.682	20
Boosting	0.734	0.724	0.756	0.694	20
Random Forest	0.711	0.704	0.724	0.685	20
Bagging (estimator=DecisionTree)	0.710	0.705	0.720	0.691	36
Boosting	0.738	0.732	0.750	0.715	36
Random Forest	0.711	0.707	0.720	0.694	36

The comparison presented in Table 4 between the MPML approach and traditional ensemble methods demonstrates clear performance advantages of MPML, particularly when expert knowledge and diverse perspectives are incorporated into the learning process. Top-performing models are indicated in bold.

The most notable results come from the MPML *(cardio_expert_and_stat_grouping)* configuration, which significantly outperforms all other methods with an accuracy of 0.852, F1 Score of 0.851, precision of 0.858, and recall of 0.852. Even the simpler MPML setup, *cardio_all_expert_grouping,* achieves 0.772 accuracy, surpassing all traditional ensemble methods, including Boosting and Random Forest with 100 base classifiers.

Traditional ensemble methods show consistent but limited improvements as the number of base classifiers increases. For example, Boosting with 100 classifiers reaches 0.738 accuracy, while reducing the number to 36 classifiers yields 0.738 accuracy, indicating a performance plateau. Similarly, Random Forest and Bagging exhibit minor variations in performance regardless of the ensemble size.

The MPML approach not only enhances predictive accuracy but also improves the balance between precision and recall, which is evident from the nearly identical values across all evaluation metrics for the best MPML configuration. These results underscore the effectiveness of MPML in producing more robust and reliable models compared to conventional ensemble methods.

Paired t-Test Comparison (with Cardiovascular Disease dataset)

The paired t-test results demonstrate that MPML significantly outperforms traditional ensemble models, including Boosting, Bagging, and Random Forest, across all key evaluation metrics: accuracy, precision, recall, and F1 score. The t-statistics for these comparisons are exceptionally high, ranging from approximately 66 to 130, with p-values consistently below 0.05, indicating that the performance improvements seen with MPML are statistically significant and extremely unlikely to be due to random chance. Importantly, the magnitude of these t-statistics far exceeds typical thresholds for significance, meaning the differences observed are not subtle but reflect strong, measurable advantages in favour of MPML. Notably, MPML achieves these superior results with only 36 base classifiers, while the competing models utilize 100 or more, highlighting MPML's efficiency.

These results emphasize that MPML delivers more reliable and balanced predictions, particularly in scenarios where both high precision and recall are essential. The most significant statistical gains are observed in recall and F1 score, where large positive t-statistics reflect MPML's ability to correctly identify more positive instances without sacrificing precision, which is crucial in domains like healthcare or fraud detection. The extremely high t-statistics across all metrics not only confirm statistical significance for individual model comparisons, but also signal that MPML's advantages are substantial and consistent across different performance dimensions.

5.4. Interpreting The MPML Model

In this section, we present a systematic approach to deconstructing and visualizing the inner workings of the MPML model in a manner that is accessible to non-technical audiences. This is achieved by extracting and interpreting feature impact scores and feature directions. The feature impact score quantifies the degree to which a specific feature influences the model's prediction for a given instance, providing insight into the feature's contribution to the decision-making process. In parallel, the feature direction indicates the directional influence of the feature, specifying toward which class the feature shifts the model's prediction. Together, these components offer a transparent, interpretable view of the model's behaviour, enhancing both understanding and trust in the system's outputs.

To support this interpretability framework, the model leverages Platt Scaling, specifically the sigmoid method, to convert raw decision scores from the decision tree classifier into calibrated probability estimates. Platt Scaling is a well-established technique for transforming the raw output scores of classification models into calibrated probability estimates (Böken, 2021), thereby enhancing both the interpretability and the reliability of the model's probabilistic predictions. This is implemented using the *CalibratedClassifierCV* class from scikit-learn with the *'sigmoid'* option, where the model is trained with 5-fold cross-validation to ensure reliable probability calibration.

For the interpretation examples presented in this study, we utilize the model developed for heart disease prediction, trained on the heart disease dataset. The interpretability analysis is conducted using perspectives, which represent groups of related features generated through the Model Importance Grouping method described in the previous section.

Local Interpretations

Table 5: Feature Impact scores for Perspective 1 - Actual Class = 1 (Prediction - 1)

Features (Removed)	Probability: Class 0	Probability: Class 1	Feature Impact Score	Feature Pull Direction
All Features	**0.2008**	**0.7992**	-	-
`chest_pain_type`	0.326	0.674	**0.1252**	Class 1
`cholesterol`	0.2041	0.7959	**0.0033**	Class 1
`max_heart_rate`	0.2061	0.7939	**0.0053**	Class 1
`oldpeak`	0.2414	0.7586	**0.0406**	Class 1

The feature impact scores presented in Table 5 illustrates how individual features influence the model's prediction for specific instances, providing critical insights into the interpretability of the MPML model. In the first example, where the actual class is 1 (presence of heart disease), the model initially predicts class 1 with a high probability of 0.7992 using Platt Scaling. Upon systematically removing features, we observe that the probability of class 1 decreases when key features like *chest_pain_type, cholesterol, max_heart_rate*, and *oldpeak* are omitted. The calculated feature impact scores confirm that each of these features contributes positively toward classifying the instance as heart disease, with *chest_pain_type* showing the most significant impact (0.1252 with Platt Scaling). The directionality indicated by the "Feature Pull Direction" column reveals that these features collectively pull the model's prediction toward class 1, reinforcing the classification of heart disease. The interpretability derived from these impact scores allows stakeholders, including non-technical audiences, to understand not only which features are influential but also how they shape the final prediction.

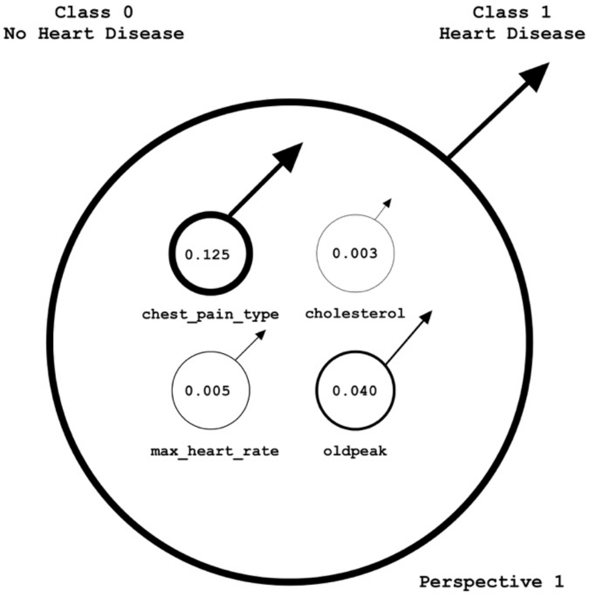

Figure 7: Feature Impact (for a single instance) and Directionality Visualization for Perspective 1

Figure 7 visually represents the **feature impact scores** and **directionality** for a specific instance within the MPML model. In this example, the outer circle labelled Perspective 1 groups the features that contributed to the model's prediction for an instance where the true class is 1 (Heart Disease). The arrows indicate the direction in which each feature influences the model's prediction.

The weight of each circle represents the magnitude of the feature impact score, while the direction of the arrow illustrates which class the feature pulls the prediction toward. The feature *chest_pain_type* has the strongest positive influence in pushing the model's prediction toward Class 1 (Heart Disease). Other features, such as oldpeak (impact score of 0.040), max_heart_rate (0.005), and cholesterol (0.003), contribute to a lesser extent but still collectively pull the prediction toward the correct class.

This visual representation enhances model interpretability by making it clear not only which features were influential but also how strongly and in which direction they affected the final classification. This could allow both technical and non-technical stakeholders to intuitively grasp the internal decision-making process of the MPML model, reinforcing confidence in the system's predictions and its ability to provide transparent, instance-level explanations for high-stakes applications like heart disease detection.

The results presented in Table 6 provide critical insights into the interpretability of the MPML model by evaluating how individual features within **Perspective 1** influence the model's prediction for a specific instance where the true class is **1 (Heart Disease)**. In this case, the model incorrectly predicted **Class 0 (No Heart Disease)** with a relatively high confidence of 81.4%. Systematically removing features reveals that each contributes to pulling the model's prediction toward Class 0, as indicated by the negative feature impact scores across all features.

Table 6: Feature Impact scores for Perspective 1 - Actual Class = 1 (Prediction - 0)

Features (Removed)	**Probability: Class 0**	**Probability: Class 1**	**Feature Impact Score**	**Feature Pull Direction**
All Features	**0.814**	**0.186**	-	-
`chest_pain_type`	0.4454	0.5546	**-0.3686**	Class 0
`cholesterol`	0.4486	0.5514	**-0.3654**	Class 0
`max_heart_rate`	0.5849	0.4151	**-0.2291**	Class 0
`oldpeak`	0.7705	0.2295	**-0.0435**	Class 0

Notably, **chest_pain_type** and **cholesterol** exert the most substantial influence, with impact scores of **-0.3686** and **-0.3654**, respectively, suggesting that these features significantly reinforced the incorrect Class 0 prediction. Similarly, **max_heart_rate** and **oldpeak** also contributed to the misclassification, though to a lesser extent. The consistent pull direction of all features toward Class 0 highlights how the model's internal representation of this instance was dominated by feature patterns associated with the absence of heart disease, leading to an erroneous outcome. This type of analysis is vital for identifying systematic biases or weaknesses in the model and informs potential avenues for feature refinement, data augmentation, or model retraining to improve prediction reliability, particularly in critical healthcare applications.

Table 7: Perspective Impact for a given instance - Actual Class = 1 (Prediction - 1)

Perspective (Removed)	**Probability: Class 0**	**Probability: Class 1**	**Perspective Impact Score**	**Perspective Pull Direction**
All Features	**0.0113**	**0.9887**	-	-
Perspective 1	0.0974	0.9026	0.0861	Class 1
Perspective 2	0.0114	0.9886	0.0001	Class 1
Perspective 3	0.0113	0.9887	0	None
Perspective 4	0.0114	0.9886	0.0001	Class 1

The results presented in Table 7 evaluate the impact of removing individual "perspectives" on the prediction probability for a given instance where the actual class is 1 and the predicted class is also 1. The baseline probability with all features included shows a strong prediction confidence for Class 1 (98.87%). When **Perspective 1** is removed, the probability for Class 1 drops significantly to 90.26%, resulting in a **Perspective Impact Score** of 0.0861, indicating that this perspective strongly supports the model's confidence in Class 1. The **Perspective Pull Direction** for Perspective 1 is towards Class 1, showing that its removal weakens the model's belief in Class 1. Conversely, the removal of **Perspective 2**, **Perspective 3**, and **Perspective 4** has negligible impact on the prediction, with minimal changes in probability (Impact Scores near 0), suggesting these perspectives contribute little to the model's confidence for this particular prediction. Notably, Perspective 3 shows no measurable impact, confirming its irrelevance in this context. Overall, the table indicates that Perspective 1 plays a significant role in supporting the prediction, while the other perspectives have little to no influence.

This local-level insight into the model's decision-making process can be highly valuable for clinicians evaluating whether the model's reasoning aligns with established medical standards and clinical judgment. By isolating the impact of individual features or "perspectives," as shown in Table 9, clinicians can assess whether the factors the model relies on to make confident predictions correspond to medically relevant indicators. Such transparency allows for critical, case-specific review, enabling clinicians to interpret whether the model is making decisions consistent with evidence-based practice. Ultimately, this process can foster either greater trust and adoption or necessary scepticism and further refinement.

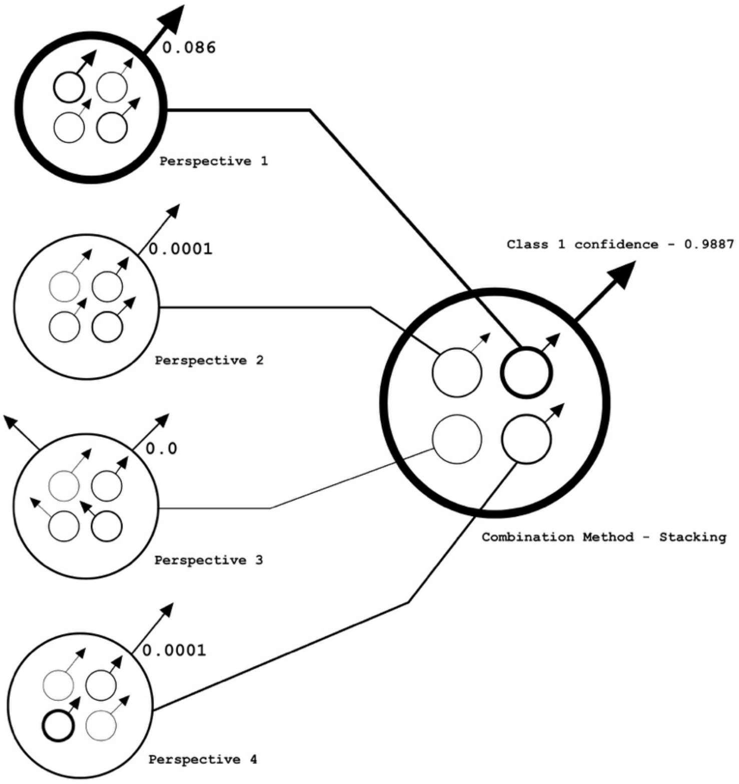

Figure 8: Perspective Contribution Breakdown for a single instance

Figure 8 presents a visual breakdown of how the MPML (Multi-Perspective Machine Learning) model combines different perspectives to arrive at a final prediction for a specific instance. The diagram shows that **Perspective 1** contributes the most to the model's confidence in predicting Class 1, with an impact score of **0.086**, while the other perspectives show minimal or no meaningful influence. Visualizations like this could help users, such as clinicians, gain insight into how the model arrives at its decision for an individual case by highlighting which perspectives the model depends on. This form of local interpretability may support users in determining whether the model's reasoning aligns with clinical expectations or established medical knowledge. This approach lays the groundwork for **global interpretations**, as systematically analysing local behaviours across multiple instances can reveal consistent patterns of perspective importance, biases, or shortcomings within the model.

Table 8: Feature Impact scores for Perspective 7 - Actual Class = 1 (Prediction - 1)

Features (Removed)	**Probability: Class 0**	**Probability: Class 1**	**Feature Impact Score**	**Feature Pull Direction**
All Features	**0.3456**	**0.6544**	-	-
`ap_hi`	0.3816	0.6184	**0.0144**	Class 1
`ap_lo`	0.3378	0.6622	**-0.0079**	Class 0
`age`	0.1729	0.8271	**-0.1727**	Class 0
`active`	0.5758	0.4242	**0.2318**	Class 1

Table 8 shows how the model's predicted probability of heart disease changes when individual features are removed. With all features included, the model already leans toward predicting heart disease for this patient, with a probability of 0.6544. Each subsequent row in the table represents the effect of removing one feature and recalculating the prediction to see how much that feature influenced the outcome. The Feature Impact Score and the Perspective Pull Direction indicate whether the presence of a given feature is pushing the model toward predicting Class 0 (no heart disease) or Class 1 (heart disease). If removing a feature increases the predicted probability of heart disease, it means that the feature was acting as a protective signal, its presence was helping the model lean toward "no heart disease." Conversely, if removing a feature lowers the predicted probability of heart disease, the feature was acting as a risk signal, contributing evidence toward a heart disease prediction.

When the systolic blood pressure feature (ap_hi) is removed from the model, the predicted probability of heart disease decreases slightly, from 0.6544 to 0.6184. The positive Feature Impact Score indicates that systolic blood pressure is pulling the model toward Class 1 (heart disease). In practical terms, this means the patient's actual systolic value provides some evidence in favour of heart disease. This aligns with well-established clinical findings: elevated systolic blood pressure is a strong, independent predictor of cardiovascular and coronary events (Palaniappan et al., 2002). Large cohort studies consistently show that systolic blood pressure is often the most important blood-pressure measure for predicting cardiovascular mortality in both untreated and treated individuals. Recent research further confirms that systolic hypertension remains a major driver of adverse cardiovascular outcomes, even after accounting for diastolic pressure and other contributing factors

(Fernández-Ruiz, 2019). Therefore, the model's interpretation in treating higher systolic BP as a risk-enhancing factor is entirely consistent with the medical literature.

When the diastolic blood pressure feature (ap_lo) is removed, the predicted probability of heart disease increases slightly, from 0.6544 to 0.6622. The negative Feature Impact Score indicates that diastolic pressure is pushing the model toward Class 0, meaning the patient's actual diastolic value acts as a weak protective signal. The effect is modest, especially compared with more influential features such as age and physical activity. This pattern aligns with the mixed findings in the cardiovascular literature: while systolic blood pressure is generally recognized as the stronger predictor of cardiovascular disease risk, particularly in older adults, diastolic pressure still has prognostic importance (Benetos et al., 2002). Both elevated diastolic pressure (as in isolated diastolic hypertension) and excessively low diastolic pressure in patients with coronary disease have been associated with adverse outcomes (Yano et al., 2022). Therefore, the model's treatment of diastolic blood pressure as having a smaller, partially protective influence for this patient is reasonable and consistent with established clinical understanding that systolic pressure typically contributes more to overall cardiovascular risk stratification than diastolic pressure.

When the age feature is removed, the model's predicted probability of heart disease increases dramatically from 0.6544 to 0.8271. This large negative Feature Impact Score indicates that age is acting as a strong protective factor for this patient. In practical terms, the model is effectively saying that because this patient is approximately 44 years old, they are less likely to have heart disease than their other risk factors alone would suggest. This interpretation makes sense given the typical age distribution of heart disease: although age is one of the strongest non-modifiable risk factors for cardiovascular disease, risk increases most steeply in older adults (Rodgers et al., 2019). Large epidemiological studies and widely used risk calculators consistently highlight age as a central driver of cardiovascular risk (Zhao et al., 2024). However, this also means that individuals who are significantly younger than the typical heart-disease population, such as this 44-year-old patient, often receive a "protective" adjustment from the model. Thus, the model's behavior aligns with clinical understanding: while age increases cardiovascular risk overall, for comparatively younger individuals in a high-risk dataset, age acts as a mitigating factor, reducing the predicted likelihood of heart disease.

Understanding that the patient is physically active (active = 1), the results reveal an important insight into how the model is interpreting this variable. With all features included, the predicted probability of heart disease is approximately 0.65. However, when the active feature is removed, the probability drops substantially to around 0.42. This means that the presence of active = 1 is increasing the model's estimate of heart disease risk. Clinically, this is counterintuitive: being physically active is widely recognized as protective against cardiovascular disease, while inactivity increases risk (Perry et al., 2023). The only reasonable interpretation is that the model has learned a dataset-specific pattern in which "active = 1" correlates with heart disease, even though this relationship does not hold physiologically. This likely reflects sampling bias, confounding, or noise in self-reported lifestyle data rather than a genuine causal link. Importantly, this example highlights the strength of the MPML framework used here: it exposes hidden or misleading associations within the model, giving users critical insight into when the model's reasoning is trustworthy and when caution is warranted.

Global Interpretations

Global interpretations are derived by aggregating the impact scores of individual features across all instances within each perspective and calculating the average contribution of each feature towards a particular class direction. Similar to local interpretations, this process is conducted separately for each perspective; however, rather than focusing on a single instance, it provides a broader overview of the general influence that each feature and perspective exert on the model's overall behaviour**.** Below, we examine the global impact score for a single feature and a single perspective. This approach enables the identification of consistent patterns, feature dependencies, or potential sources of bias at the global level, offering insights into the model's alignment with domain-specific knowledge and its potential reliability in real-world applications.

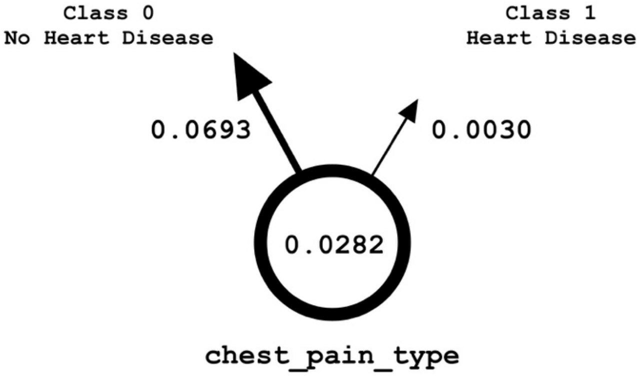

Figure 9: Global impact score of a single feature

Figure 9 provides a global interpretation of the MPML model's behaviour by illustrating the average influence of the feature *chest_pain_type* across all predictions in the dataset. The central value (0.0282) represents the mean impact score of the feature, quantifying its overall contribution to the model's decision-making process. Arrows extend from the central node to indicate the direction and magnitude of this feature's influence toward each class: 0.0693 toward Class 0 (No Heart Disease) and 0.0030 toward Class 1 (Heart Disease).

This visualization reveals that *chest_pain_type* contributes more strongly to predictions of the absence of heart disease than to its presence. The thickness and directionality of the arrows help identify how the model leans when interpreting this feature. Such global insights are critical for validating whether the model's logic aligns with clinical understanding. If the model's weighting of chest pain types reflects known medical risk factors, its use in decision support may be justified. Conversely, disproportionately low or high influence toward either class could signal underlying bias or overfitting, warranting further analysis. As such, visual tools like this support interpretability, transparency, and trust in clinical ML applications.

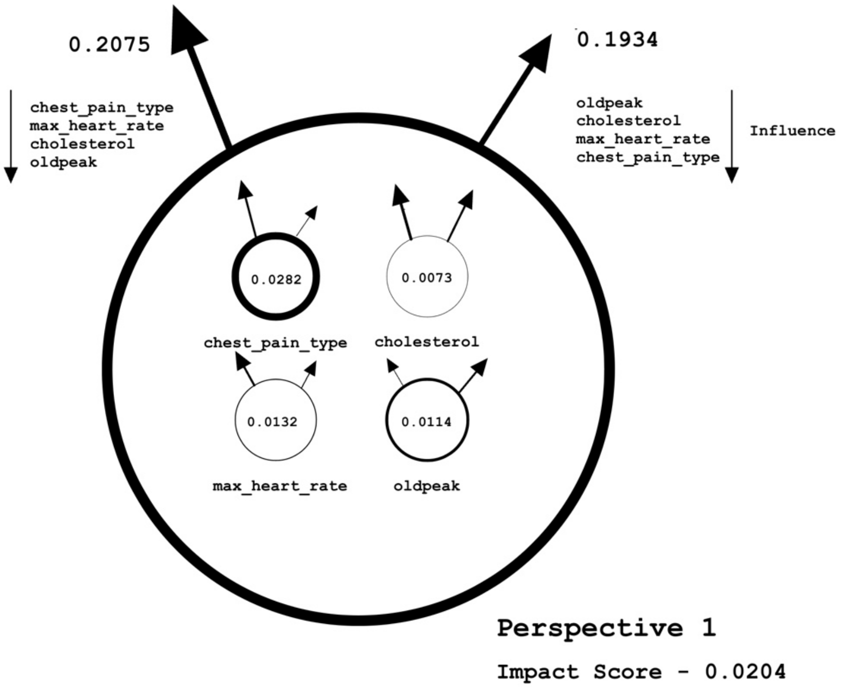

Figure 10: Global impact of a single Perspective

Figure 10 provides a global interpretation of the internal behaviour of Perspective 1, highlighting how feature-level contributions within a single perspective influence the model's overall decision-making process. The large outer circle represents the aggregated behaviour of the perspective, with an overall impact score of 0.0204. The two large arrows extending from the perspective indicate its average (across all instances) directional influence toward each class: 0.2075 toward Class 0 (No Heart Disease) and 0.1934 toward Class 1 (Heart Disease).

Inside the perspective, individual features—*chest_pain_type, cholesterol, max_heart_rate,* and *oldpeak*—are shown with their own average impact scores. Among these, *chest_pain_type* (0.0282) exhibits the highest influence, followed by *max_heart_rate* (0.0132), *oldpeak* (0.0114), and *cholesterol* (0.0073). Each feature also has directional arrows indicating whether its contribution leans more toward predicting heart disease or not.

This type of visualization can help users evaluate whether the model's learned importance for each feature aligns with clinical reasoning. For instance, chest pain type being the most influential factor supports known medical insights, whereas lower scores for cholesterol and oldpeak might invite further scrutiny. If unexpected patterns are observed—such as medically irrelevant features dominating predictions—it may highlight potential sources of bias. Overall, such perspective-level views enhance transparency and can guide validation, trust, and refinement of the model in clinical settings.

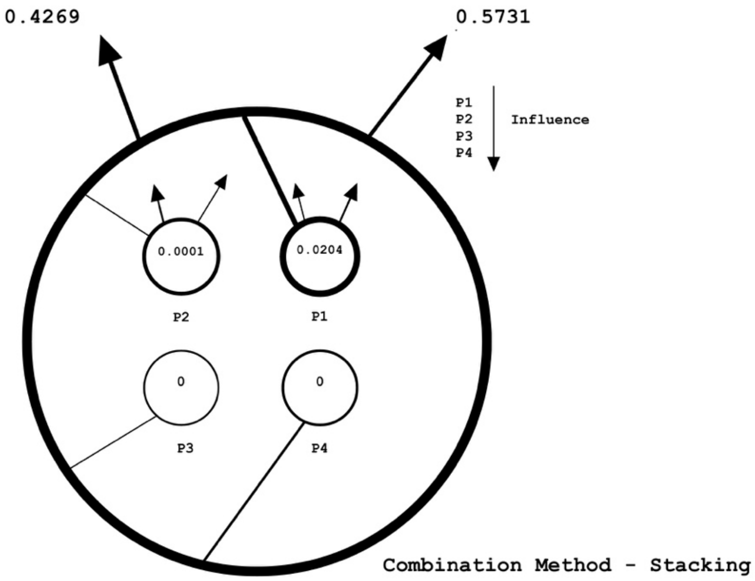

Figure 11: Global impact of all Perspectives on each class

Figure 11 presents a global summary of how each perspective (P1 to P4) contributes to the final prediction within the MPML ensemble when using the stacking combination method. The large outer circle represents the ensemble-level decision space, with the arrows indicating the directional influence toward each class: 0.4269 toward Class 0 (No Heart Disease) and 0.5731 toward Class 1 (Heart Disease).

Inside the ensemble, each sub-circle represents a specific perspective. Perspective 1 (P1) demonstrates the highest impact score (0.0204), indicating it is the most influential contributor to the final prediction. Perspective 2 (P2) exerts a minimal influence (0.0001), while Perspectives 3 and 4 (P3 and P4) show no measurable impact in this instance (0.0000), suggesting their contribution to the ensemble's final decision was negligible.

This visualization enables users to understand not only which perspectives are active but also how much they shape the model's outcome. The arrows illustrate how these perspectives influence the predicted class directionally, reinforcing the interpretability of the ensemble structure. If high-impact perspectives, like P1, are based on medically meaningful features, this can affirm the clinical validity of the model. However, the inactivity of P3 and P4 could either reflect redundancy or insufficient signal, which may warrant further investigation.

Overall, such ensemble-level explanations provide transparency into how stacked predictions are constructed, making it easier to verify whether the ensemble relies on robust, clinically grounded insights—or if adjustments to grouping, weighting, or architecture are needed before deployment in sensitive domains like healthcare.

6. Discussion and Limitations

6.1. Discussion

The results from both the Heart Disease and Cardiovascular Disease datasets provide compelling evidence of the superiority of the Multi-Perspective Machine Learning (MPML) framework over traditional machine learning and ensemble methods. Across both datasets, MPML consistently delivers higher predictive performance while maintaining a more compact model structure, which is particularly advantageous for real-world applications such as healthcare, where computational efficiency and interpretability are essential.

In the Heart Disease dataset, conventional classifiers such as Naive Bayes, Decision Trees, and SVM demonstrated limited predictive power, with SVM yielding notably poor results across all metrics. While standard ensemble methods like Bagging, Boosting, and Random Forest significantly outperformed these baselines, MPML achieved comparable or superior results with fewer base classifiers. For example, MPML with expert grouping using only 7 base classifiers achieved an accuracy of 0.92475, outperforming Boosting with 100 classifiers. The advanced MPML configurations that combine mutual information, correlation analysis, PCA, model importance, and expert knowledge further improved performance, achieving an exceptional accuracy and F1 score of 0.997 using 23 base classifiers. In contrast, traditional ensemble methods required up to 1000 base classifiers to approach, but not match, this level of performance.

The paired t-test comparisons reinforced these findings, demonstrating that MPML's performance advantages are not only consistent but also statistically significant. Across accuracy, precision, recall, and F1 score, MPML significantly outperformed Boosting, Bagging, and Random Forest, with p-values at or near zero and t-statistics as high as 18.12. These results underscore MPML's ability to deliver both high performance and efficiency, offering more reliable predictions with fewer computational resources.

A similar pattern emerged with the Cardiovascular Disease dataset, where traditional ensemble methods achieved modest performance improvements with increased base classifiers, yet plateaued well below MPML's best configurations. Notably, the MPML configuration that combined statistical feature grouping with expert-driven grouping achieved an accuracy of 0.852 and similarly high precision, recall, and F1 scores, substantially outperforming all competing models, including those with significantly higher ensemble sizes. Even the simpler MPML setup outperformed Bagging, Boosting, and Random Forest, reinforcing the value of integrating diverse perspectives, including domain expertise, into the learning process.

The paired t-test results on this dataset provided even stronger statistical evidence of MPML's superiority. With t-statistics exceeding 100 for accuracy comparisons and similarly large values for precision, recall, and F1 score, the differences were not only statistically significant but also practically substantial. These findings illustrate MPML's robustness and its ability to generalize effectively across different datasets and problem domains.

Within Perspective 1 key features such as chest pain type, cholesterol, oldpeak, and maximum heart rate emerged as highly influential across all instances. This aligns with their established significance in cardiovascular risk assessment. Chest pain type has been consistently identified as a critical diagnostic indicator for heart disease, differentiating between typical angina, atypical angina, and non-anginal pain patterns associated with ischemic events (Végh et al., 2024). Elevated serum cholesterol levels are well-documented contributors to atherosclerosis and coronary artery disease, directly impacting predictive models' ability to assess risk (Logan et al., 2020; Logan et al., 2024; R. Raja, 2025). Similarly, oldpeak, which measures ST-segment depression induced by exercise relative to rest, provides crucial information about myocardial ischemia and has been recognized as a robust predictor of cardiovascular outcomes in stress test evaluations (Savchuk & Doroshenko, 2025). Maximum heart rate achieved during exercise testing reflects cardiac reserve capacity and is strongly correlated with cardiovascular health and disease risk (Islam et al., 2024). The prominence of these features within MPML underscores its capacity to prioritize clinically relevant parameters, reinforcing both its predictive validity and potential for clinical adoption. This could strengthen trust in AI-assisted decision support systems.

A key advantage of the MPML framework lies in its ability to maintain both high predictive performance and interpretability. Unlike traditional ensemble methods, which often operate as black boxes, MPML incorporates mechanisms for interpreting model behaviour, such as feature importance rankings derived from its diverse grouping strategies. This combination of transparency and predictive strength makes MPML well-suited for sensitive domains like healthcare, where understanding model outputs is critical for building trust and ensuring responsible decision-making. The interpretability provided by MPML not only aids in model validation but also allows clinicians and domain experts to trace predictions back to relevant features and perspectives, aligning machine learning outputs with human expertise.

To further support responsible AI deployment in healthcare, it is essential to consider how MPML addresses concerns related to bias, patient consent, and fairness in decision-making. Machine learning models trained on clinical data are susceptible to biases that stem from imbalanced datasets, underrepresentation of subpopulations, or systemic disparities in care. MPML mitigates these risks by allowing feature groupings to be informed by domain knowledge, enabling models to be audited not only globally but also at the perspective level. This makes it possible to assess whether certain demographic or clinical subgroups are disproportionately influencing predictions or receiving skewed outcomes. Furthermore, MPML's layered interpretability enables transparent communication of how and why a specific decision was made, facilitating better-informed discussions with patients and healthcare providers. This transparency supports the ethical imperative of informed patient consent, where individuals must understand how automated tools influence their care. By clearly attributing predictions to specific, meaningful feature groups (e.g., lab results, symptoms, demographics), MPML enhances accountability and fairness, reducing the risk of opaque or unjust recommendations and aligning machine learning predictions with the principles of equitable, patient-centered care.

MPML is particularly well-suited to domains where interpretability is critical and where rich domain knowledge already exists, such as cardiovascular medicine with its extensive risk scores and clinical guidelines. In such settings, the upfront cost of expert-guided perspective construction is justified by the resulting transparency and alignment with clinical practice.

6.2. Limitations

Despite its demonstrated strengths, the MPML framework is not without limitations. One notable drawback is the overhead associated with setting up the various perspectives that underpin the model's multi-faceted design. Unlike traditional ensemble methods such as Bagging or Random Forest, which automatically generate diverse feature subsets or data samples, MPML requires a deliberate and often time-consuming process to group features into predefined categories based on domain knowledge or statistical criteria. This setup phase introduces additional complexity and may slow down deployment, particularly in scenarios where expert input is limited or unavailable.

Another limitation of MPML relates to the computational cost of obtaining impact scores for global interpretation. Generating these interpretations requires running the model iteratively for the number of features involved, which can be computationally expensive, especially for datasets with a large number of features. While the global interpretations provide valuable insights into feature importance and model behaviour, they are effectively static unless the model is retrained. Consequently, if new data becomes available or if the feature space evolves, the interpretation process must be repeated, further adding to the computational demands.

These limitations imply that while MPML offers significant performance and interpretability advantages, its adoption may be constrained by resource availability and the need for expert-driven feature grouping. Future work should explore automating aspects of the perspective setup process and optimizing the computational efficiency of impact score calculations to broaden the framework's accessibility and scalability.

A formal prospective evaluation, such as user-centred studies with clinicians or deployment within a live clinical workflow, was beyond the scope of this study, but remains essential for establishing the real-world utility and practical impact of the MPML framework.

7. Conclusion and Future Work

7.1. Conclusion

The Multi-Perspective Machine Learning (MPML) model proposed in this study has demonstrated clear advantages over traditional classifiers and ensemble methods across multiple datasets. MPML consistently outperformed standard models such as Bagging, Boosting, and Random Forest in both predictive accuracy and overall evaluation metrics, even when using significantly fewer base classifiers. This efficiency, combined with superior performance, highlights MPML's potential as a reliable and scalable solution for complex classification tasks, particularly in healthcare.

Beyond raw performance, MPML also addresses a critical gap in conventional ensemble methods by providing interpretable impact scores that reveal the relative influence of individual features on model predictions. This transparency is particularly valuable in medical applications, where clinician trust and alignment with domain knowledge are essential. The integration of expert-driven feature grouping, statistical perspectives, and dimensionality reduction allows MPML to deliver not only high predictive accuracy but also meaningful, interpretable outputs that align with clinical reasoning.

However, while MPML succeeds in enhancing both performance and interpretability, several important limitations remain that warrant attention in future work. Most notably, the computation of global impact scores currently requires multiple model runs, which may become computationally expensive for large or high-dimensional datasets. Because these scores are generated from a fixed training distribution, the resulting global interpretations are also static and may become outdated as underlying data distributions shift over time. Future research should therefore explore more efficient and adaptive techniques for generating global interpretability outputs, as well as methods to streamline the perspective setup process and ensure that MPML continues to capture the full complexity of clinical scenarios. Overall, the findings of this study position MPML as a high-performing, interpretable, and efficient ensemble framework with strong potential for deployment in sensitive, high-stakes domains such as healthcare.

A key limitation of this study is that we did not perform a formal evaluation of interpretability with clinicians or other end-users, nor did we deploy MPML in a real clinical workflow. The qualitative expert review we report provides only preliminary support for clinical plausibility. Future work should therefore include controlled user studies and prospective evaluations that measure the impact of MPML's explanations on clinical decision-making, workload, and trust.

7.2. Future Work

Future research will also focus on validating the MPML model in real-world clinical settings. Although the model has shown promise in experimental conditions, integrating it into clinical workflows is essential. By collaborating with healthcare professionals, we will gather feedback to refine the model, ensuring it enhances decision-making and patient outcomes in practice. Furthermore, to assess the generalizability of the MPML approach, we will apply the framework to different healthcare datasets. This will test the model's robustness and accuracy across diverse clinical domains, identifying potential limitations and ensuring it remains effective in various settings.

Another critical area of focus is enhancing the completeness of the MPML model. This involves ensuring that all relevant factors and interactions are captured, providing a more comprehensive view of the decision-making process for clinicians. By refining the model to account for complex relationships within the data, we aim to provide healthcare professionals with a more reliable tool for clinical decision support. These future directions will significantly enhance the practical utility, reliability, and interpretability of the MPML framework, bringing us closer to making AI-driven healthcare systems both transparent and trustworthy.

A primary focus for the future development of this work will be the establishment of a robust and reliable metric specifically designed for evaluating interpretable ensemble models. This metric would extend beyond traditional machine learning evaluation criteria to include measures of interpretability and comprehensiveness, particularly tailored to healthcare applications.

A valuable direction for future work is to systematically evaluate MPML's training and inference times in comparison to standard ensemble models such as Bagging, Boosting, and Random Forest. While MPML offers enhanced interpretability through its multi-perspective structure, its computational demands—particularly due to training multiple sub-models and aggregating their outputs—may impact its suitability for real-time clinical applications.

Future studies should benchmark MPML against traditional ensembles using diverse clinical datasets to assess scalability, latency, and computational overhead under practical deployment scenarios. In particular, exploring optimizations for inference, such as model pruning, parallelization, or selective perspective invocation, could enhance MPML's viability for time-sensitive tasks.

Additionally, the feasibility of modular updating and batch inference should be examined in dynamic clinical settings where data evolves and decisions are not always time-critical. Such evaluations will provide clearer guidance on when and how MPML can be deployed effectively in clinical decision-support systems.

A key limitation of MPML in its current form is the reliance on domain expertise for perspective construction. Defining clinically meaningful feature groups requires input from clinicians or other domain experts, which introduces additional effort and may limit scalability to settings where such expertise is scarce. This design choice was intentional, as it grounds perspectives in clinically interpretable constructs, but it also means that fully automated deployment is not yet possible.

Future work should explore integrating or comparing MPML with automated feature-grouping and AutoML frameworks, such as NiaAML, to reduce the manual effort required for perspective construction and enhance scalability across domains with limited expert availability.

References

Acharya, D., B, D., & Nair, R. P. (2025). *Explainable AI in Healthcare: A Stacking-Based Approach for Diabetes Classification*. 1–6. https://doi.org/10.1109/ICSSES64899.2025.11009637

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aditya, P. S. R., & Pal, M. (2022). *Local Interpretable Model Agnostic Shap Explanations for machine learning models*.

Al-bakri, F. H., Bejuri, W. M. Y. W., Al-Andoli, M. N., Ikram, R. R. R., Khor, H. M., & Tahir, Z. (2025). A Meta-Learning-Based Ensemble Model for Explainable Alzheimer's Disease Diagnosis. *Diagnostics*, *15*(13), 1642–1642. https://doi.org/10.3390/diagnostics15131642

Ali, Z. (2025). *Heart Disease Prediction Using AI Models: A Comparative Study on the Sulianova Dataset*. https://doi.org/10.5281/ZENODO.15337802

Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Panahiazar, M., Koohestani, A., Khozeimeh, F., Nahavandi, S., & Sarrafzadegan, N. (2019). A database for using machine learning and data mining techniques for coronary artery disease diagnosis. *Scientific Data*, *6*(1), 227–227. https://doi.org/10.1038/s41597-019-0206-3

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 310–310. https://doi.org/10.1186/s12911-020-01332-6

Awe, O. O., Mwangi, P. N., Goudoungou, S. K., Esho, R. V., & Oyejide, O. S. (2025). Explainable AI for enhanced accuracy in malaria diagnosis using ensemble machine learning models. *BMC Medical Informatics and Decision Making*, *25*(1), 162–162. https://doi.org/10.1186/s12911-025-02874-3

Bassan, S., Amir, G., Zehavi, M., & Katz, G. (2025). *What makes an Ensemble (Un) Interpretable?* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.08216

Benetos, A., Thomas, F., Bean, K., Gautier, S., Smulyan, H., & Guize, L. (2002). *Prognostic value of systolic and diastolic blood pressure in treated hypertensive men*. Archives of internal medicine, 162(5), 577-581.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Bühlmann, P. (2012). Bagging, Boosting and Ensemble Methods. In *Handbook of Computational Statistics* (pp. 985–1022). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21551-3_33

Dieber, J., & Kirrane, S. (2020). *Why model why? Assessing the strengths and limitations of LIME*.

Fernández-Ruiz, I. (2019). *Systolic and diastolic hypertension independently predict CVD risk*. Nature Reviews Cardiology, 16(10), 578-579.

Gulati, S., Guleria, K., & Goyal, N. (2022, April). *Classification and detection of coronary heart disease using machine learning.* In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1728-1732). IEEE.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting Black-Box Models: A

Review on Explainable Artificial Intelligence. *Cognitive Computation*, *16*(1), 45–74. https://doi.org/10.1007/s12559-023-10179-8

Islam, A., Shanto, M. N. I., Dipto, T. R., Rabby, Md. S. M., & Monna, H. F. (2024). *Classifying Heart Diseases: An Ensemble Technique Combining With Federated Learning*. 1–6. https://doi.org/10.1109/ICRPSET64863.2024.10955879

Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018). Artificial Intelligence in Cardiology. *Journal of the American College of Cardiology*, *71*(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

Li, M., Sun, H., Huang, Y., & Chen, H. (2024). Shapley value: From cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, *4*(1), 2–2. https://doi.org/10.1007/s43684-023-00060-8

Logan, K., Asemota, H., Nwokocha, C., A Lawrence, M., Thompson, R., Nwokocha, M., & Bakir, M. (2020). The Effects of Synthesized Semicarbazone Copper Complex on Blood Pressure in Normotensive and L-NAME Induced Hypertensive Rats. *Journal of Biotechnology and Biomedicine*, *03*(02). https://doi.org/10.26502/jbb.2642-91280028

Logan, K., Nwokocha, C., Asemota, H., & Gray, W. (2024). Characterization of ACE inhibitory activity in Dioscorea alata cv and its implication as a natural antihypertensive extract. *Journal of Ethnopharmacology*, *319*, 117221–117221. https://doi.org/10.1016/j.jep.2023.117221

Mahajan, A., Sharma, N., Aparicio-Obregon, S., Alyami, H., Alharbi, A., Anand, D., ... & Goyal, N. (2022). A novel stacking-based deterministic ensemble model for infectious disease prediction. Mathematics, 10(10), 1714.

Manu Siddhartha. (2025). *Heart Disease Dataset (Comprehensive)*.

Merrick, L., & Taly, A. (2020). *The Explanation Game: Explaining Machine Learning Models Using Shapley Values* (pp. 17–38). https://doi.org/10.1007/978-3-030-57321-8_2

Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Miller, S. T., & Busby-Earle, C. (2017). *Multi-perspective Machine Learning (MPML)—A Machine Learning Model for Multi-faceted Learning Problems*. 363–368.

Palaniappan, L., Simons, L. A., Simons, J., Friedlander, Y., & McCallum, J. (2002). *Comparison of usefulness of systolic, diastolic, and mean blood pressure and pulse pressure as predictors of cardiovascular death in patients≥ 60 years of age (The Dubbo Study)*. The American journal of cardiology, 90(12), 1398.

Panda, M., & Mahanta, S. R. (2023). *Explainable artificial intelligence for Healthcare applications using Random Forest Classifier with LIME and SHAP*.

Panhalkar, A. R., & Doye, D. D. (2022). A novel approach to build accurate and diverse decision tree forest. *Evolutionary Intelligence*, *15*(1), 439–453. https://doi.org/10.1007/s12065-020-00519-0

Perry, A. S., Dooley, E. E., Master, H., Spartano, N. L., Brittain, E. L., & Pettee Gabriel, K. (2023). *Physical activity over the lifecourse and cardiovascular disease*. Circulation research, 132(12), 1725-1740.

R. Raja. (2025). Integrating Machine Learning Approaches for Predictive Analysis of Heart Disease Risk Factors. *Communications on Applied Nonlinear Analysis*, *32*(8s), 456–468. https://doi.org/10.52783/cana.v32.3690

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *" Why should i trust you?" Explaining the predictions of any classifier*. 1135–1144.

Robnik-Sikonja, M., & Kononenko, I. (2008). Explaining Classifications For Individual Instances. *IEEE Transactions on Knowledge and Data Engineering*, *20*(5), 589–600. https://doi.org/10.1109/TKDE.2007.190734

Rodgers, J. L., Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L. E., Shah, K., ... & Panguluri, S. K. (2019). *Cardiovascular risks associated with gender and aging*. Journal of cardiovascular development and disease, 6(2), 19.

Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: Application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, *34*(10), 1013–1026. https://doi.org/10.1007/s10822-020-00314-0

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–206.

Saini, A., & Prasad, R. (2022). *Select Wisely and Explain: Active Learning and Probabilistic Local Post-hoc Explainability*. 599–608. https://doi.org/10.1145/3514094.3534191

Salih, A., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Menegaz, G., & Lekadir, K. (2024). *A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME*. https://doi.org/10.1002/aisy.202400304

Sapra, L., Sandhu, J. K., & Goyal, N. Intelligent Method for Detection of Coronary Artery Disease with Ensemble Approach, 2021.

Saridena, A., Saridena, A., & Kethar, J. (2023). A Supervised Deep Learning Model for the Detection of Cardiovascular Disease. *Journal of Student Research*, *12*(4). https://doi.org/10.47611/jsrhs.v12i4.5178

Savchuk, D., & Doroshenko, A. (2025). Explainable AI methods to increase trustworthiness in healthcare. In *Responsible and Explainable Artificial Intelligence in Healthcare* (pp. 55–89). Elsevier. https://doi.org/10.1016/B978-0-443-24788-0.00003-0

Topuz, K., Bajaj, A., Coussement, K., & Urban, T. L. (2025). Interpretable machine learning and explainable artificial intelligence. *Annals of Operations Research*, *347*(2), 775–782. https://doi.org/10.1007/s10479-025-06577-w

Végh, A., Takáč, L., Czakóová, O., Dancsa, K., & Nagy, D. (2024). Comparative Analysis of Machine Learning Classification Models in Predicting Cardiovascular Disease. *International Journal of Advanced Natural Sciences and Engineering Researches*, *8*(6), 23–31.

Zhao, D., Wang, Y., Wong, N. D., & Wang, J. A. (2024). *Impact of aging on cardiovascular diseases: from chronological observation to biological insights: JACC family series*. JACC: Asia, 4(5), 345-358.

Zhao, J., Xie, X., Xu, X., & Sun, S. (2017). Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, *38*, 43–54. https://doi.org/10.1016/j.inffus.2017.02.007