INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Enhancing Supervised Learning Robustness Investigating the Impact of Label Noise on Algorithm Performance

ANA CATARINA FARIA RUA
Junho de 2025

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

Enhancing Supervised Learning Robustness

Investigating the Impact of Label Noise on Algorithm Performance

Ana Catarina Faria Rua

Dissertation submitted in fulfilment of the requirements for the degree of Master of Informatics Engineering, Specialization Area in Data Science

Supervisor: Doutora Maria de Fátima Coutinho Rodrigues

Porto, June 2025

Statement of Integrity

I hereby declare that I have conducted this academic work with integrity.

I have not plagiarized or applied any form of undue use of information or falsification of results along the process leading to its elaboration.

Therefore, the work presented in this document is original and authored by me, having not previously been used for any other end. The exceptions are recognized in the section “Ethical considerations” of the first chapter. This section also states how AI tools were used and for what purpose.

I further declare that I have fully acknowledge the Code of Ethical Conduct of P.PORTO.

ISEP, Porto, 29 de June de 2025

Dedicatory

“However difficult life may seem, there is always something you can do and succeed at”- Stephen Hawking

Abstract

Supervised learning serves as the foundation for many AI systems because it enables models to learn from labelled examples. However, label noise resulting from human annotation errors or systematic biases can diminish model performance and limit generalization capabilities. This challenge is particularly significant in critical domains such as healthcare, finance, and autonomous systems. This thesis focuses on studying the impact of label noise on supervised learning algorithms in order to evaluate its influence across different datasets and to propose robust strategies for mitigation. This project includes methods of loss correction, data augmentation, and advanced noise detection frameworks as examples and demonstrates their prospective advantages through empirical experiments. The provided noise-robust algorithms in the research can be used with any real-world scenarios to improve the resilience of the algorithm. The findings are meant to be a connection between academic research and practical implementation by offering guidelines for handling noisy datasets effectively while ensuring model reliability and fairness. The proposed approach raised the average F1-Score from 0.647 under baseline conditions to 0.757 after full optimization.

Keywords: Supervised learning, label noise, machine learning, noise correction, ensemble learning, anomaly detection

Resumo

A aprendizagem supervisionada constitui um dos pilares fundamentais dos sistemas de inteligência artificial, permitindo, assim, que os modelos sejam treinados com base em exemplos rotulados. No entanto, o ruído nos rótulos, originado por erros humanos durante o processo de anotação ou por vieses sistemáticos, compromete de forma significativa o desempenho dos modelos limitando as suas capacidades de generalização. Este problema é particularmente relevante em domínios críticos como a saúde, as finanças e os sistemas autónomos, onde decisões incorretas podem ter consequências graves. Esta dissertação tem como objetivo investigar o impacto do ruído nos rótulos em algoritmos de aprendizagem supervisionada, avaliando a sua influência em diversos conjuntos de dados e propondo estratégias robustas para a sua mitigação. O trabalho contempla o desenvolvimento e a aplicação de métodos avançados, incluindo funções de perda adaptadas ao ruído, estratégias de aumento de dados e frameworks de deteção de ruído, cujas vantagens são demonstradas por meio de análises teóricas e de experiências empíricas. A abordagem proposta subiu os valores de F1-Score de 0.647 para 0.757 após otimizações.

Palavras-Chave: Aprendizagem supervisionada, ruído nos rótulos, robustez, correção de ruído, aprendizagem automática

Acknowledgment

I would like to express my deepest gratitude to those who supported me throughout the development of this thesis and also the completion of my Master's Degree.

To my partner in life, Daniel Lourenço, for all of his unconditional love and patience, and for being by my side through every step of this journey. For the constant encouragement and all the help, and specially for never letting me give up during this path.

To my parents, Anabela and Paulo, and my sister and brother in law, Sofia and Jorge, for their endless patience, love, and belief in me, even in the most challenging moments. This encouragement has been the foundation of my perseverance.

I would specially like to extend my sincere gratitude to my supervisor, Professor Fátima Rodrigues, whose guidance, availability, and constructive feedback were vital during this process. This commitment and willingness to help were fundamental to the successful completion of this work.

To Rui Mendonça, my colleague and friend throughout the Master's Degree, who collaborated with me in our many joint projects, and most of all, was my company in all classes from the very beginning.

I would extend my gratitude to my team at work, in particularly Ana, Nuno and Alexandre, who encouraged and supported me during this demand period.

Finally, I dedicate a heartfelt thank you to those who are no longer with us, my grandparents, but whose memory continues to inspire and accompany me during life.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ML	Machine Learning
SVM	Support Vector Machine
RF	Random Forest
NB	Naïve Bayes
GNB	Gaussian Naïve Bayes
LOF	Local Outlier Factor
IF	Isolation Forest
CRISP-DM	Cross-Industry Standard Process for Data Mining
LLM	Large Language Model
GDPR	General Data Protection Regulation

1 Introduction

1.1 Contextualization

The widespread use of machine learning has attracted the public's interest to its underlying machine learning models that make performance predictions and decisions based on vast amounts of labelled data. Of course, these programs are going to be very dependent on the quality of their databases in all such industries [1].

More often than not, real-world data is imperfect because it contains noisy labels injected into it as a result of the annotation process. Label noise, which could be both random and systematic, constitutes a threat to unbiased models, reduced accuracy, overfitting, and consequently to the robustness of machine learning systems.

Machine learning is utilized across a wide range of industries and for various applications, with different sectors exhibiting varying rates of adoption and implementation. While some sectors prioritize AI for routine tasks, others focus on decision support rather than immediate action [2]. Many companies do deploy algorithms in urgent scenarios where rapid human response and effective problem analysis are essential. Therefore, addressing the potential effects of label noise in critical settings is central to the success of these organizations. Best practices for mitigating label noise are not only fundamental for maintaining high performance but also play a crucial role in ensuring fairness and ethical coding. Therefore, considering the potential effects of label noise in critical settings and how to solve it is central to the success of companies in the future. Best practices for the label noise elimination problem are not just covered as a fundamental requirement for the preservation of high performance but also, they are the base for fairness and ethical coding [3].

The report's main goal is to provide an all-encompassing solution to the main challenges posed by label noise in supervised learning systems, which will be the quantification of the extent of noise in labelled datasets and the methodical identification of new methods for its alleviation. This research can be seen as an attempt to enhance the robustness, fairness, and ethical integrity of machine learning models deployed in high-stakes applications by exploring new

corrective measures for label noise, including the in-depth study of advanced correction and model adaptation methods.

1.2 Problem Statement

Label noise, which can be caused by human annotation errors or systematic biases in the data labelling process, poses a significant challenge in supervised learning frameworks, particularly in high-stakes domains such as healthcare, finance, and autonomous systems [4]. The presence of inaccurate labels can lead to poor model performance, undermining the ability of algorithms to generalize effectively to unseen data. As organizations increasingly rely on AI systems for critical decision-making, the implications of label noise become more pronounced, potentially resulting in erroneous outcomes that could affect human lives, financial stability, and operational efficiency. Despite the apparent significance of this issue, many prevailing machine learning models do not adequately account for the impact of label noise, resulting in models that may fail to perform as intended in real-world applications.

Addressing label noise is not only crucial for enhancing the robustness and accuracy of supervised learning models; it also underpins the ethical deployment of AI technologies. Inaccurate labelling can perpetuate biases and inequalities, raising concerns about fairness and accountability in AI systems. Thus, the challenge of label noise necessitates the development of effective strategies for noise detection, mitigation, and robust model training [5]. The purpose of this thesis is to explore implications of label noise for model performance and fairness, and examine contemporary methods for addressing these challenges within supervised learning paradigms.

1.3 Objectives

Based on the challenges outlined before, this project is designed to examine the systematic label noise in supervised learning and to elaborate a framework capable to detect and mitigate this label noise. By addressing both practical evaluation and methodological development, the study seeks to strengthen the reliability of predictive models in high-stakes domains.

Therefore, the present study aims to inject noise on different levels into a heterogeneous collection of datasets, and quantify the resulting degradation in detection metrics, such as precision, recall and F1-score. It then analyzes the influence of class imbalance, feature dimensionality and decision-boundary overlap on noise detection. To do so it is proposed to develop a hybrid pipeline that combines a calibrated ensemble of classifiers, k-Means++ clustering and anomaly detectors and optimize confidence threshold and contamination rates through systematic hyperparameter search that enhance the F1-Score. It is also necessary to measure the gains in terms of accuracy obtained by retraining a model on clean data and then address if noise filtering induces information loss. Finally establish a practical guideline for threshold selection, dataset sizing and reproducible experimentation in critical application scenarios.

1.4 Methodology Selection

The methodology used for the research of this study has been adapted to fit the needs proposed in the problem statement and objectives section. Accordingly, the use of Cross Industry Standard Process for Data Mining (CRISP-DM) methodology admits an autonomous documentation for each phase allowing the generation of a comprehensive study of the discoveries made on this theme.

Since the CRISP-DM approach allows the documentation of the phases, it is possible to have a life cycle of the developments proposed well conducted, like stated by [6], that believed in the use of this methodology as a managing form of the report and evaluation of the objects generated from one phase to the other one.

The model has an iterative recursion within. The first phase has the key insights and prerequisites identified during the state of the art, that leads to the second phase, data understanding, in which the heterogeneous sources are explored and structed. The third phase the data representations are harmonized in order to preserve their integrity. The next phase, the modeling, is responsible for building the proposed pipeline and then assessing its results on the evaluation phase. The last phase is not applicable to this study but in this phase all prior activities are consolidated and deployed. A full explanation of each phase is detailed in the chapter three.

1.5 Research Questions

According to the project objectives, the investigation of the impact of label noise on algorithm performance is structed around a group of research questions designed to guide the systematic evaluation of label noise detection and mitigation approaches. The questions edge both the experimental methodology and the analysis of practical findings, as demonstrated in table 1.

Table 1- Research Questions

Identifier	Research Questions
RQ1	How effective is the proposed hybrid pipeline in detecting and mitigating label noise across diverse datasets, compared to existing baseline approaches?
RQ2	How do dataset characteristics, such as class imbalance, feature dimensionality, and decision boundary overlap, influence the pipeline's detection performance and robustness?
RQ3	What impact does noise correction have on downstream predictive accuracy, and under which conditions does aggressive filtering compromise model generalization?

First, an investigation on how different noise levels application can impact precision, recall and F1-score across different datasets. Then, the examination of the role played by specific dataset characteristics in modulating noise detection. After this characterization, the study will do a comparison between the proposed hybrid pipeline, integrating a calibrated ensemble classifier, and the conventional baseline approaches. In addition, a systematic hyperparameter search will be conducted in order to identify which hyperparameter maximize the F1-score. Moreover, the investigation will quantify the gains in predictive accuracy obtained by retraining models on clean data and will identify scenarios in which aggressive noise filtering leads to loss of information. Finally, practical guidelines for threshold selection, dataset sizing and reproducibility will be formulate providing recommendations for critical application contexts.

1.6 Artificial Intelligence Tools

During the writing of this thesis, the Grammarly plugin was used to enhance the quality of the written work. Grammarly is one of such platforms, that employs artificial intelligence technologies to help authors correct mistakes in grammar, spelling, and style or provide recommendations for improving the text in regards to clarity, coherence, and formality [7].

With regard to linguistics issues of the thesis, the use of Grammarly was restricted to the mechanical aspects of the work. In particular, the application was used to check for the likelihood of errors relating to grammar and spelling, checked for the consistency of key terms and phrases, and finally the content's formal tone and readability. The application was also used to provide suggestions on how to enhance clarity and coherence of the content in terms of meeting scholarly standards.

All the suggestions developed by Grammarly were subject to an evaluation and all changes to the text were made only when it felt necessary. The program was aimed at providing technical support only, with no relationship whatsoever with the generation of ideas relating to the study, interpretation of results and the conclusions which were made in this thesis. It was only supportive help maintain good language and stylistic quality.

In addition, during the experimental development phase, a large language model (LLM), was used as a programming assistant to support certain technical tasks. This included refining code implementation, debugging complex functions, and clarifying technical documentation or existing methods referenced in the literature. The use of this tool was strictly limited to supporting the independent work, primarily by providing guidance for technical implementation details and assisting with code structuring. Importantly, all scientific reasoning, experimental design, methodology formulation, interpretation of results, and conclusions drawn in this thesis are the sole intellectual independent contribution.

By acknowledging the application of such tools, this section reinforces the commitment to maintaining the highest standards of academic integrity while using supportive technologies responsibly.

1.7 Document Structure

The document is organized in a manner that is intended to first identify the problems of label noise in supervised learning systems and then the solutions.

The first chapter gives the overall idea of the thesis, the importance of label noise as a challenge in machine learning, the statement of the problem, research objectives and the research questions to be addressed in the study.

The second chapter is devoted to the analysis of the existing literature on the problem and represents the development of the supervised learning approach as well as increasing attention paid to the label noise issue. This chapter also provides a comprehensive analysis of the existing state-of-the-art approaches to deal with noise, the limitations of these works and the theoretical base upon which this work is developed.

The third chapter is dedicated to the methodology where it is described the research approach used and also the evaluation criteria taking in consideration.

The fourth chapter gives a summary of the datasets used and the techniques implemented during the experimental design.

The fifth chapter evaluates the performance of the iterations implemented and discuss the results making comparisons and taking conclusions analyzing the results.

Finally, the sixth chapter gives the overall conclusion of the thesis, the major conclusions of the research as well as the recommendations for future research.

This structure was design to make sure that there is a clear and consistent evolution of the thinking from the definition of the problem to the creation of the solution and its assessment, in line with the objectives of the study.

2 State of the Art

2.1 Introduction

Label noise is a growing concern in supervised learning tasks, as noisy examples may severely cripple the training of models and result in skewed or erroneous predictions [5]. As data volume continues to grow and the practice of outsourcing annotations to human labelers or automated machines gains traction, the reality of label noise becomes inevitable [1]. Such a problem is particularly heinous in domains such as healthcare, finance, and high-stakes decision-making scenarios, where erroneous instances may fan out ethical, legal, and operational misconduct of even greater proportions than raw technical competence [8]. In light of this fact, the current work examines approaches for systematic identification and diminution of label noise, placing focus on the incorporation of clustering and heterogeneous ensemble methods.

The growing body of literature in label noise has led to numerous approaches attempting to enhance robustness and reverse its effect. These methods are the noise-robust loss functions, probabilistic models, confidence filtering, ensemble learning, and cluster-based correction methods[5] . However, all of the proposed approaches make either a random noise distribution assumption or leverage noise structure prior knowledge, limiting their applicability in real cases where noise patterns are likely to be complex, structured, and difficult to estimate [5]. Besides, using multiple complementary detection mechanisms in aggregation is also not explored, particularly for systematic noise [5].

This work applies a heterogeneous ensemble-based method to detect and fight against systematic label noise. It draws conceptual insights from the model proposed by [9] and further develops it by incorporating multiple unsupervised anomaly detection models in a multi-stage noise filtering system. The approach presented here relies on the fusion of supervised ensemble confidence, cluster consistency, and feature-label consistency employing Isolation Forest and Local Outlier Factor models. The fusion allows both low-confidence instance identification and class-inconsistent instance detection based on structural features of the feature space and thereby addressing various sides of systematic label noise.

2.2 Label Noise in Supervised Learning

Label noise is referred to as errors in class labels of training instances [4]. Although most supervised learning algorithms presume clean and properly labelled data sets as a fundamental assumption, this assumption does not usually hold in practice. Label noise can be divided into two classes: random noise and systematic noise [10], [11].

Random noise tends to be feature-independent, appearing as uniformly distributed stochastic labelling errors across the dataset. Systematic noise, by contrast, is highly correlated with particular feature subsets or class boundaries, leading to systematic misclassification patterns that are particularly difficult to disentangle from genuine class structures [4]. Systematic label noise is far more troublesome because it has a tendency to introduce biases that are correlated with true decision boundaries, potentially reinforcing harmful spurious correlations [12].

Several studies have reported the adverse effects of label noise on model performance, particularly in healthcare [13], fraud detection [14], and autonomous systems [15], where even low rates of label errors can cause significant deterioration in predictive accuracy, generalization capacity, and fairness. Additionally, the presence of label noise has also been shown to exacerbate model overfitting, variance, and distort feature importance estimates [16].

Classic supervised approaches such as decision trees, support vector machines, and neural networks each have varying sensitivities to label noise depending on their inherent flexibility, regularization strategies, and loss functions. Deep neural networks, as high-capacity models, have the propensity to memorize noise in training data and require special regularization or correction techniques to preserve generalization performance when there is noise [17].

2.3 Existing Approaches for Noise Mitigation

The minimization of label noise has been addressed by a range of methods, which can be categorized into loss correction, noise filtering, semi-supervised learning, and ensemble methods.

Loss correction methods attempt to correct the loss function in-place to counteract noise, often through noise transition matrices or probabilistic transformations [18][19]. While they work well in some circumstances, they typically need to know the noise distribution, or at least have a good estimate of it, something that is hardly ever possible in practice. They also lose their effectiveness when the noise distribution exhibits rich, instance-specific behavior, as it usually does in systematic noise conditions.

Noise filtering methods attempt to identify and remove potentially mislabeled samples prior to model training. Confidence score-based, k-nearest neighbor consistency-based, or clustering label agreement-based methods have been widely investigated [20]. Confidence filtering calculates the uncertainty of the classifier for every instance and flags low-confidence instances as potential noisy [21]. Single confidence scores may be unreliable in high-dimensional feature spaces, particularly with noisy examples along the class boundaries.

Ensemble learning has exhibited strong robustness to label noise through prediction aggregation from a number of learners [22]. Heterogeneous ensemble learning, where

classifiers are of different bias and sensitivity, can achieve enhanced robustness by heightened difficulty in all models perpetuating noise-inducing errors simultaneously. In particular, calibration of classifiers so as to generate well-calibrated probability estimates supports more trusted confidence-based filtering [23].

The approach made by [24] demonstrates the strength of integrating supervised ensemble confidence and unsupervised cluster consistency. Through the fusion of multiple weak noise indicators as well as precision. Their two-stage approach—ensemble confidence filtering followed by K-Means cluster consistency, provides the conceptual foundation for this work.

Unsupervised anomaly detection models, such as Isolation Forest [11] and Local Outlier Factor [25], have also been found helpful in identifying feature-label inconsistencies. These methods make use of structural deviation in the feature space to identify samples whose features are not consistent with their predicted labels [10], [11] . While more typically used for unsupervised outlier detection, their class-conditional application to label noise detection remains a research area of active interest [5], [26].

Finally, semi-supervised approaches have been explored for greater robustness to label noise by using unlabeled instances or generating pseudo-labels [27]. These are highly sensitive to the quality of the original labels, and can amplify mistakes when significant noise already exists.

2.4 Systematic Label Noise: Key Research Gaps

Despite significant progress in noise-robust learning, several critical gaps persist regarding systematic label noise.

In the first place, most existing works continue to focus on random label noise, while systematic noise is less investigated due to its complexity. As systematic noise is generally strongly correlated with feature subspaces or subpopulations, it has the effect of reinforcing hidden biases, magnifying ethical concerns of fairness and accountability [28]. This is particularly troublesome for high-stakes applications where systematic mislabeling disproportionately affects vulnerable populations.

Second, the majority of existing solutions assume knowledge of the noise structure or noise transition probabilities [18]. In practice, such knowledge is rarely available, especially when noise is generated by complex domain-specific biases or annotation inconsistencies. Robust, adaptive solutions that do not rely on prior noise distribution knowledge are becoming ever more necessary.

Third, while ensemble methods have been encouraging, most existing ensemble methods utilize homogeneous ensembles, ensembling multiple instances of the same kind of classifier. There is little exploration in the literature of heterogeneous ensemble models that take advantage of disparate classifier biases to enhance robustness in the presence of systematic noise, like [24] have demonstrated.

Fourth, there is limited exploration of the integration of supervised and unsupervised approaches into unified pipelines that can cross-validate noisy label detection based on multiple independent signals. Although confidence-based filtering and clustering label

consistency both offer partial solutions, their intersection is under-utilized, even as it has the potential to improve detection precision and recall [5], [27].

Finally, there is a shortage of practical, reproducible frameworks that distil these intricate methodologies into real-world pipelines for application to heterogeneous datasets. The majority of existing studies are confined to small-scale, domain-specific datasets, which limits wider generalizability [29].

The present work attempts to fill these gaps by designing and systematically evaluating a hybrid multi-phase detection and mitigation system that combines supervised ensemble confidence filtering, clustering consistency, via K-Means++, and unsupervised anomaly detection, Isolation Forest and LOF, over a number of publicly available datasets from diverse domains.

2.5 Conceptual Framework for the Present Work

The proposed work posits a multi-step noise detection and mitigation pipeline, in which systematic label noise is identified and minimized by sequentially concatenating heterogeneous detectors. The approach borrows from the two-step model discussed in [24] but builds upon it through iterative updates, hyperparameter tuning, and other anomaly detection models.

In the first stage, a variety of supervised classifiers, Calibrated SVM, Naive Bayes and Random Forest, are trained and calibrated probability scores are used to estimate per-sample confidence levels. Low-confidence ensemble samples are labelled as tentative candidates for noisy labels.

In the second phase, several unsupervised techniques are used to validate or correct the initial filtering. K-Means++ clustering is utilized to project cluster assignments on class labels, marking instances of cluster-class inconsistency. Concurrently, Isolation Forest and Local Outlier Factor are run class-conditioned to identify feature-label inconsistencies not identified by the supervised models.

Several combination strategies are studied to merge these signals, including conservative consensus filtering, majority voting, and a strict Two-Stage filter, keeping only samples identified by both stages as confirmed noise candidates. This combination provides a flexible trade-off between high precision, not removing clean data, and high recall, complete detection of noise, which is crucial for effective systematic noise removal.

Finally, after noise detection, there is a noise mitigation step where flagged samples are removed from the training set and models are retrained with the cleaned datasets to assess performance improvement. Comparison with baseline models allows quantitative comparison of the usefulness of the proposed approach across various datasets and noise levels.

2.6 Summary

The problem of label noise continues to be one of the main hindrances to dependable machine learning, especially in domains with complex, structured noise patterns. While much of the earlier work has proposed a multitude of solutions to noise mitigation, most current solutions

are still limited either by their assumption of random noise, by computational complexity, or by diminished generality to a diversity of domains.

This thesis contributes towards addressing these limitations through the presentation of a novel, heterogeneous multi-stage approach that integrates calibrated ensemble learning, unsupervised clustering consistency, and anomaly detection techniques for structured label noise detection. The approach extends and builds on top of the early ideas of [24] for applicability on diverse datasets and for quantitatively assessing detection performance, mitigation effectiveness, and resulting improvements in model robustness.

3 Methodology

This chapter describes in detail the methodological approach employed during this research. The research was developed following the CRISP-DM methodology guidelines, which offer an organized and methodical process for data analysis and machine learning projects. Adhering to CRISP-DM allowed for a methodical going through steps such as problem understanding, data preparation, modelling, testing and interpretation [9].

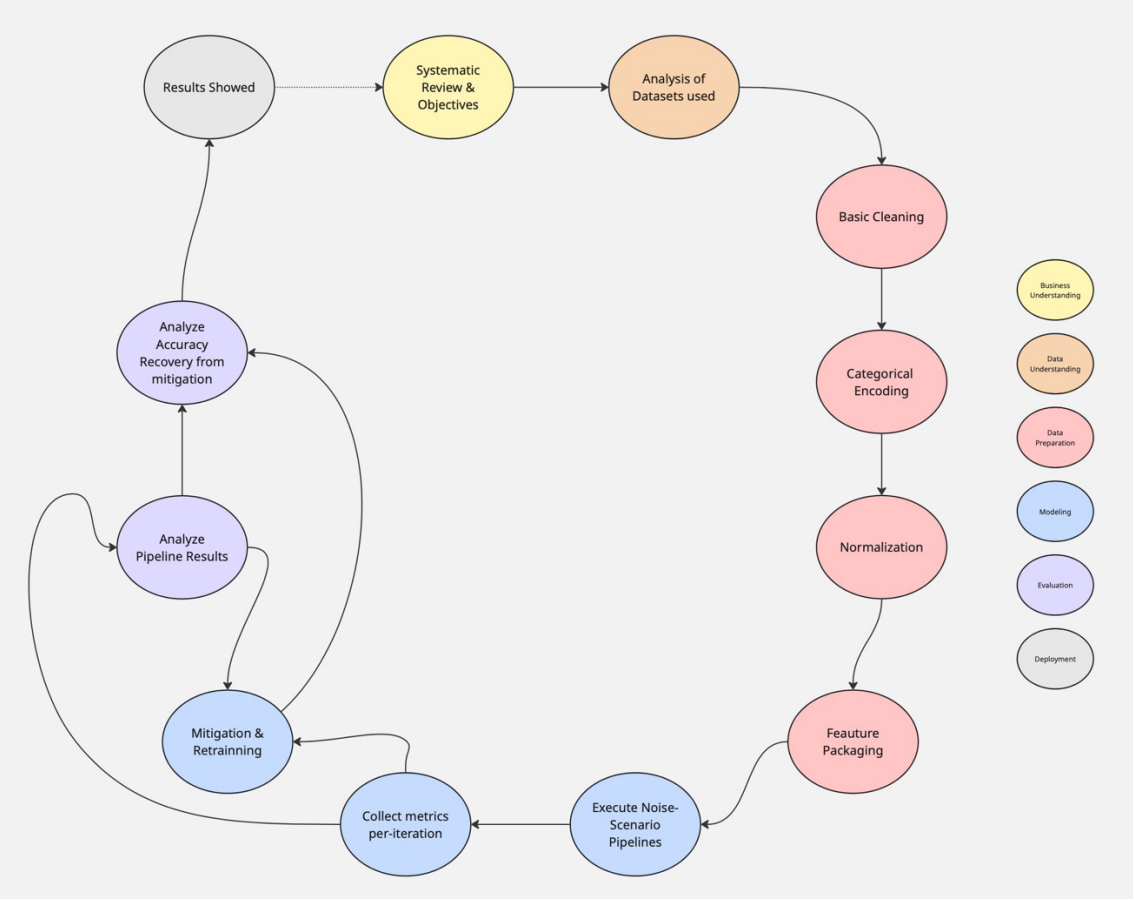

Figure 1- Methodology Diagram Based on CRISP-DM adapted to the study

Figure 1 gives an overview on the methodology used during this study, following an adaptation of the CRISP-DM model approach to structure the implementations done. Further there is a detailed explanation on each phase represented on the figure.

3.1 Problem Understanding

The increasing reliance on supervised learning algorithms in domains where data labelling is prone to human error or uncertainty has put tremendous focus on the problem of label noise. In particular, systematic label noise, where consistent patterns of mislabeling affect specific subsets of the data, can severely compromise model generalization and reliability. This issue is particularly pressing in sensitive application areas such as healthcare and finance, where decisions based on models can have drastic consequences.

The main research aim of the project was to investigate and develop practical noise detection and mitigation strategies that can identify systematically mislabeled instances and improve the downstream performance of supervised models. Rather than seeking solely very specialized or domain-specific solutions, the project sought to propose broadly applicable detection frameworks that leverage supervised and unsupervised learning principles.

3.2 Data Understanding and Acquisition

To avoid the imposed limitations on generalizability of the suggested methods, a diverse collection of datasets was collected from well-known public databases, primarily the UCI Machine Learning Repository and Kaggle. The datasets were chosen to span across various domains, including healthcare, finance, education, and general-purpose classification tasks, thereby introducing heterogeneity in the size of the data, class imbalance, feature types, and complexity of the classification boundaries.

Every dataset was carefully inspected for data quality, class distribution confirmation, and their availability for noise injection and evaluation experiments. Such initial inspection was critical to guarantee that the artificially injected noise would be under systematic control and that the measures of evaluation would reflect effective differences in performance among datasets and techniques.

3.3 Data Preparation

Following data gathering, a thorough preprocessing pipeline was implemented to ensure consistency and comparability across datasets. This process comprised multiple carefully design stages, intended to prevent the introduction of spurious variability and to prepare the data for the requirements of noise detection algorithms.

The initial step was the deletion of identifier or index columns that contained no predictive information. Categorical features were transformed with Label Encoding to generate numerical forms from discrete nominal variables for application in supervised and unsupervised techniques. Numerical features were also tested for missing values, which were filled with mean

substitution to ensure the datasets' integrity. Following imputation, all the features were normalized using standard scaling, meaning zero mean and unit variance, to prevent excessive weightage to any single feature due to differences in scales. This scaling operation was necessary, particularly for the unsupervised methods, such as clustering and outlier detection methods, that are highly sensitive to feature scales. Lastly, the target features were converted to consistent numerical labels, allowing for equal treatment of the output space across data sets.

This preprocessing pipeline was used uniformly on all data sets so that the following analysis was performed under uniform and controlled circumstances.

3.4 Modelling

The experimental approach was structured around three development iteration stages of algorithmic evolution, where each stage was developed based on the limitations and findings identified in the previous phase.

3.4.1 Iteration one – baseline Confidence-Based Ensemble

In the first experimental stage, a supervised ensemble confidence model was designed as the baseline noise detection mechanism. This approach fused three various base classifiers, Support Vector Machines, Gaussian Naïve Baye, and Random Forest, each of which was calibrated using isotonic regression to produce well-calibrated probabilistic predictions. The predicted class probabilities of the calibrated classifiers were averaged for each instance, which resulted in a combined confidence score representing the ensemble's confidence in the provided noisy label.

Examples with low aggregated confidence were flagged as suspicious noisy samples on the hypothesis that true mislabeled examples will create uncertainty between multiple diverse classifiers. 0.5 was used as the threshold to split low-confidence examples. This ensemble approach was complemented with a set of unsupervised detectors, such as K-Means++ clustering, Isolation Forest, and Local Outlier Factor. The cluster-based detector relied on cluster-to-class consistency, while the outlier detectors leveraged feature-label consistency within each class.

The first iteration provided a full baseline by taking into account single detectors as well as their logical combinations, which enabled comparison of complementary strengths between supervised confidence estimation and unsupervised anomaly detection.

3.4.2 Iteration two – Hyperparameter Optimization

The second iteration aimed to address shortcomings observed in the static threshold structure of the initial ensemble model. To that end, an extensive hyperparameter tuning mechanism

was included for optimizing the detection capabilities of the ensemble and unsupervised models.

For the ensemble detector, a grid search over possible confidence thresholds was performed to identify the value that resulted in the optimal F1-score on the noisy training folds. Similarly, the contamination parameters of the Isolation Forest and LOF algorithms were swept iteratively over a pre-specified range, with the aim of making precision-recall trade-offs optimal for detection. The K-Means++ method was held fixed, as a baseline unsupervised method throughout iterations.

This iteration eliminated the combination of detectors and focused instead on enhancing the individual performance of each method through careful tuning. The introduction of optimized hyperparameters demonstrated measurable improvements in detection sensitivity and precision when compared to the static configuration of the first iteration.

3.4.3 Iteration three – Combined Ensemble and Two-Staged Cascading Approach

In the third and final iteration, the insights gained in the previous stages were consolidated into a hybrid multi-stage detection framework. Here, the optimized ensemble detector and unsupervised anomaly detection were merged more integrally.

The ensemble model was first optimized to discover an optimal low-confidence threshold, following the same procedure as in iteration two. Concurrently, the Isolation Forest contamination rate was re-optimized to determine its optimal setting. With these optimized pieces in place, a Two-Stage cascade methodology was used: first, the samples that were identified as low-confidence by the ensemble were further filtered by the unsupervised detectors. Only the samples marked by both the ensemble and at least one of the unsupervised methods, Isolation Forest or K-Means++, were ultimately marked as noisy.

Furthermore, majority voting systems were explored, consolidating signals from all detectors to provide additional robustness against false positives and false negatives. This ensemble of methods allowed cross-validation of noise signals from multiple independent perspectives, supervised, clustering-based, and anomaly-driven, resulting in a highly robust and adaptive noise detection system.

3.5 Evaluation Protocol

The testing process also followed a carefully controlled experimental procedure to ensure objectivity and comparability. Synthetic label noise was added to the training partitions of every dataset according to a systematic noise model, with rates of contamination being ten, twenty and thirty percent. The noise injection procedure ensured class-balanced mislabeling, which simulates real-world scenarios where systematic misclassification affects all classes proportionally.

The performance of each detection technique was measured by comparing the detected noisy instances to the known set of injected noisy labels. Precision, recall, and F1-score were reported

for each technique at each level of noise, providing a comprehensive quantitative analysis of detection accuracy.

Apart from detection accuracy, a cause evaluation was conducted to quantify the effect of noise mitigation on downstream supervised learning accuracy. Identified noisy examples were removed from the training set, on which new Random Forest classifiers were retrained on the cleaned training data and evaluated on the untainted test sets. This final evaluation estimated the classification accuracy gain from the proposed noise detection and mitigation pipeline.

3.6 Ethical and Practical Considerations

While developing and testing this methodology, a series of measures designed to satisfy methodological transparency, reproducibility, and ethical responsibility were implemented. All datasets used in this study are publicly accessed on UCI Machine Learning Repository [30], Open ML [31] and Sklearn [32]. The previous contain no personal information. There were no desire to identify individuals or to link information across different sources. This approach is aligned with the requirements of the EU General Data Protection Regulation (GDPR), which demands the use of anonymized data and fully prohibits the process of information that may compromise individual privacy [33].

To consider the impact of label noise in a controlled manner preventing the introduction of uncontrolled biases, incorrect labels were injected with fixed rates, using random sampling and hard-coded seeds. This method eliminates any human intervention in noise generation and guarantees fully reproducible results by the applying the exact same seed values and noise parameters [34].

Also, the noise injection was unequivocally design to be class-agnostic, avoiding the historical biases that often manifest as systematic label errors. The application of identical noise rates on all classes prevents uneven impact on any subgroup, following the best practices for bias mitigation in machine learning [35].

This study is in conformity with the European Commission's Ethics Guidelines for Trustworthy AI, which emphasize human oversight, privacy, transparency and accountability [36]. The unemployed use of sensitive personal data, by documenting the data sources and procedures, and by validating the model performance under different noise conditions, this study supports the ethical measures and sustenance the development of reliable AI systems.

In Summary, the methodological choices employed, ensure that this study advances technical solutions for label noise and exemplifies responsible and ethical research practices.

4 Data Analysis and Experimental Development

This chapter details the data analysis and development followed during this project to evaluate the impact of label noise on supervised learning algorithms. This section documents a detailed explanation of dataset selection, preprocessing, noise injection procedures, noise detection methods, hyperparameter tuning and a possible noise mitigation approach.

4.1 Datasets Description

The experimental data analysis is performed on twelve benchmark datasets, each one selected based on diversity in complexity, size, dimensionality and varied domains. These are standard in machine learning research projects and are used frequently for benchmarking algorithms in different approaches.

Table 2- Summary of datasets used

Dataset	Instances	Features	Classes	Domain
Balance Scale[37]	625	4	3	Psychology
Breast Cancer[38]	569	30	2	Medical Diagnosis
Car Evaluation[39]	1728	6	4	Decision-Making
Dermatology[40]	366	34	6	Dermatological Diagnosis
Glass[41]	214	9	6	Forensic Identification
Iris[42]	150	4	3	Botanical Classification
Phishing[43]	11055	30	2	Cybersecurity
Satimage[31]	6435	36	6	Remote Sensing
Seeds[44]	210	7	3	Agricultural
User Knowledge[45]	403	5	4	Educational Assessment
Vertebral Column[46]	310	6	3	Orthopedics
Wine[47]	178	13	3	Chemical Analysis

As described in table 2, the first dataset, Balance Scale, belongs to the psychology domain, where balance-scale states are classified based on weight distribution. It is divided in four

numerical attributes. Its multi class structure with balanced distribution allows the study of label noise effects on a smaller but structed dataset.

The Breast Cancer dataset is provided by medical diagnosis procedures and is widely used for binary classification tasks. Its instances are described by thirty numerical features that capture various properties of breast masses.

The Car Evaluation aggregates crucial information for decision making tasks. This dataset is described by six categorical features like price, maintenance cost, safety and comfort. With this four classes, this dataset test the robustness of algorithms in multi class, categorical scenarios.

Then, the Dermatology is a specialized medical dataset with the purpose to diagnose dermatological diseases. It is organized in thirty four numerical features describing clinical and histopathological attributes.

The Glass dataset includes forensic data that is aimed to identifying types of glass from chemical properties. It is described by nine numerical features and is specially challenging due to the overlapping feature distributions across its six classes.

The Iris dataset is a popular benchmark dataset used in botanical studies, consisting of measurements of iris flowers. It is described by four numerical features such as petal length, petal width, sepal length and sepal width. Although its simplicity, this dataset provides a useful baseline scenario for studying the effects of label noise.

After, the Phishing dataset is sourced from cybersecurity contexts and contains website characteristics that allows to detect phishing attempts from legitimate websites. With thirty binary and numerical features this dataset allows a comprehensive evaluation of algorithm robustness in a large scale binary classification contexts.

The Satimage dataset is a product of remote sensing satellite imagery. Described by thirty-six numerical attributes, is used to classify land cover types. This dataset represents a scenario with a large volume of data and numerous classes, ideal for testing algorithm performance under significant complexity and dimensionality.

The Seeds dataset is sourced from agricultural contexts and consists in classifying three different wheat varieties based on seven numerical attributes connected to kernel morphology.

This dataset is essential in a typical scenario in agricultural analysis, emphasizing the relevance of accurate labelling for quality control.

The User Knowledge is related to educational assessments, divided by five numerical attributes. It categorizes user knowledge levels, simulating real-world scenarios.

The Vertebral Column dataset is used in orthopedic diagnosis and described by six numerical features representing biomechanical attributes. The goal is to classify patients based on spinal abnormalities, making it valuable in hectare diagnosis scenarios.

At last, the Wine dataset, used in chemical analysis contexts, has 13 chemical properties features, facilitating the classification of wines from different cultivars. This dataset provides an excellent scenario for examining how label noise affects chemical analysis based datasets.

The comprehensive variety and context of all those datasets allows robust experimentation and meaningful insights into the effects of systematic label noise on algorithm performance across diverse real-world scenarios.

4.2 Experimental Development and Methodological Framework

The experimental development approach is conduct in distinct phases, designed to first introduce and detect the label noise, evaluate and tune detection methodologies and, at last, perform noise mitigation. This process starts with i) phase 1, data preprocessing and noise injection; ii) evaluation and parameter tunning of label noise detection methods; iii) ensemble based label noise detection (two-staged approach) and iv) label noise mitigation and performance evaluation.

This process is iterative, so each phase is built upon the previous one, allowing to verify insights gained at each step to enhance robustness and accurately measure the impact of label noise on classifier performance.

4.2.1 Data Preprocessing

The data preprocessing is a crucial step to assure homogeneous scaled and suitable data for further use. All datasets are going through a similar detailed preprocess to achieve the optimal data status.

Firstly, and since all datasets contain identification columns that don’t have any meaningful value for the models, they are removed automatically to avoid unnecessary delays in the model’s learning process.

Subsequently, the target variable, Y, is separated from the rest of the variables, X. The first one is defined as initial parameter and is emphasized from the rest of the group, in order to turn the application of the models easier. To tolerate full compatibility with the implemented methods, all categorical variables are turned into numerical ones. This step is vital since the methods being used, like Random Forest, or SVM, require numerical representation of the data.

This transformation is performed with the help of the LabelEncoder, that converts each different category in a numerical sequential representation, assuring consistency and simplicity.

The existence of missing values in numerical variables is a very common scenario, that may jeopardize the efficiency of models. For that reason an imputation of values is made generated by the arithmetic mean of each column. This choice allows a simple approach to minimize the impact of those missing values.

After this, and because the target variable is categorical by nature, it is necessary to numerical codify it using the LabelEncoder. This transformation smooths the models process and also future evaluation of the results.

Finally, the data is normalized with the help of StandardScaler, allowing a zero mean and unit standard deviation to all the attributes. This process is especially important in sensitive scale methods, such as SVM and clustering techniques.

Although standard preprocessing operations were applied uniformly across all datasets, the aim was not to optimize preprocessing itself, but to ensure data compatibility and stability for the subsequent application of label noise detection algorithms.

4.2.2 Noise Injection Mechanism

It is necessary to artificially inject label noise in a controlled way, in order to evaluate the performance of the noise detection methodologies being implemented in this study. Since the original datasets are assumed to be clean, the controlled introduction of label noise allows the creation of a ground truth that can be used for the evaluation of the distinct detection methods.

The process of the noise injection is done by a custom made function, This allows the introduction of noise in a controlled and stratified manner, ensuring that noise is distributed proportionally across all classes of the dataset. The reason for opting for a balanced noise insertion poses on the need to prevent bias towards majority and minority classes, which could misrepresent the evaluation of the methods under test.

For each class present in the label vector y, the function will calculate the number of samples that must be corrupted according to the desired percentage of noise, in this case it is ten, twenty and thirty percent. The number of noisy instances per class are a result of the multiplication between the number of samples in the dataset and the defined noise percentage. Then, for each class, the number of instances are randomly selected without replacement. The selected instances have their true labels replaced by a randomly chosen one, excluding the original label, to guarantee label flipping. The indexes of the modified instances are stored and

returned alongside the noisy label vector, enabling a direct comparison between the ground truth of noisy samples and the detections performed by the various algorithms.

This approach allows the simulation of systematic label noise with controlled intensity, enabling the evaluation of the models' robustness under varying noise levels.

4.2.3 Auxiliary Functions for Evaluation Metrics

To assess the quality of the noise detection methods, it is necessary to define objective evaluation metrics based on the known set of injected noisy labels. The evaluation is performed using standard metrics of precision, recall and F1-score, normally used in classification problems were imbalance between true positives and false positives can affect the performance interpretation.

Two functions are created and implemented to compute these metrics based on the detected noisy instances provided by the noise injection. The first one is for calculating the precision and is defined as the proportion of correctly detected noisy instances, true positives, among all instances flagged as noisy, true positives and false positives. The second metric calculates the recall. This represents the proportion of actual noisy instances correctly identified, true positives, relative to the total number of noisy instances introduced, true positives and false negatives. The F1-score is computed as the harmonic mean of the precision and recall, providing a single measure that penalizes both false positives and negatives.

The metrics allow the systematic and quantitative evaluation of the different noise detection algorithms across the various datasets and noise levels.

4.2.4 Supervised Ensemble Classifier with Probabilistic Calibration

One of the most important core components of the noise detection approach relies on supervised learning, in particular a probabilistic ensemble classifier. This is created by combining the outputs of three distinct supervised classifiers, Support Vector Machine, Random Forest and Naïve Bayes.

The idea of combining multiple classifiers happens due to the diversity of their inductive biases and decision boundaries. When aggregating their predictions, the ensemble can leverage complementary strengths and mitigate individual weaknesses, leading to a more robust estimation when classifying potentially noisy instances.

Nevertheless, it is essential that the individual classifiers produce well calibrated probability estimates, since the goal is to quantify the confidence of the assigned labels for noise detection purposes. Poorly calibrated probabilities may lead to overconfident predictions that not reflect true classifications and can potentially degrade noise detection performance.

For this reason, all classifiers were bound using the CalibratedClassifierCV from the scikit-learn library, using isotonic regression as the calibration method. This isotonic calibration allows flexibility and superior performance when sufficient calibration data is available. A three-fold

cross-validation approach is used to fit the calibrators, ensuring that the process of calibration does not overfit the training data.

This ensemble calibration framework allowed for the computation of confident scores per instance by averaging the calibrated probabilities assigned to the modified label across all three classifiers. These confident scores serve as the basis for the noise detection mechanisms in the ensemble learning phase of the study.

4.2.5 Unsupervised Clustering via K-Means++

In addition to the supervised ensemble learning, an unsupervised clustering strategy is built to provide complementary information for noise detection. The aim is to, upon clean data, instances that belong to the same class should exhibit feature similarity and form coherent clusters in feature space. Consequently, inconsistencies between cluster membership and the assign labels may serve as an indicator of label noise.

The k-Means++ algorithm is selected for clustering due to its simplicity, scalability and improved initialization compared to the standard K-Means. For each dataset the number of clusters is set equal to the number of classes, assuming that a clean dataset would show cluster structures broadly aligned with the class distribution.

A post processing step was necessary to align the unsupervised cluster assignments with the known class labels. This is possible by searching for the permutation of cluster labels that maximized the classification accuracy when comparing the mapped cluster assignments with the true class labels. This process acknowledges an interpretation of clustering results with impact, in the context of label noise.

Once the best cluster to class mapping is established, instances whose noisy label differed from their cluster derived class are flagged as potentially noisy, thus generating a binary detection mask for evaluation purposes.

4.2.6 Experimental Evaluation and Model Development

Following the preprocessing process and initial model definitions previously explained, the goal of this study consists in a progressive and iterative framework for label noise detection evaluation. This methodological design is divided in three consecutive evaluation iterations. Each iteration build on the insights and limitations observed in the previous one. The initial iteration begins with a baseline configuration, deliberately avoiding using any hyperparameter optimization to capture the natural behavior of each noise detection method. In the next iteration, hyperparameter tuning is introduced to maximize the potential of each individual model, while still maintaining isolated evaluation. The final iteration evaluates the effect of combining these optimized detectors into integrated hybrid architectures. Throughout all

iterations, identical noise injection, evaluation metrics and dataset partitions are preserved, to ensure experimental consistency. His evolution across iterations is represented in table 3.

Table 3- Summary Table: Evolution Across Iterations

Iteration	Key Characteristics	Tunning Applied	Methods Combined	Goal
1	Default parameters	No	Yes	Establish baseline
2	Individual hyperparameter tuning	Yes	No	Isolate tunning effect
3	Tunned parameters combined	Yes	Yes	Maximize detection power

4.2.6.1 Iteration one – Baseline Ensemble and Combinatorial Approaches

In the first iteration, the performance of various noise detection methods is evaluated using default parameter settings. This stage's purpose is to serve as a baseline assessment, aiming to reveal the natural sensitivity of each technique without any form of tunning.

Different detection approaches are considered. The supervised method is based on an ensemble of three distinct classifier, SVM, RF and GNB. To enable reliable confidence estimation, each classifier is individually calibrated using the isotonic regression through a 3 fold cross-validation scheme. Calibration is required because ensemble based detection make the assumption of averaging classifier predicted probabilities.

For each example, ensemble confidence is the average of each classifier's confidence that the example's (possibly noisy) assigned label was correct. Whenever this confidence is below a chosen threshold of 0.5, the example is marked as suspected noise. The rationale for this confidence-based ensemble technique is that systematic label noise would introduce uncertainty and disagreement among the heterogeneous classifiers, so lower consensus confidence scores would ensue.

Aided by the supervised detection, several unsupervised methods are employed. K-Means++ clustering is directly applied to the scaled feature space with the number of clusters set to the number of classes. Cluster labels have the true class labels projected onto the cluster assignments by optimizing matching accuracy of labels and clusters using permutation alignment. A sample is labelled as potentially noisy if its noisy label is not consistent with the most common label in its assigned cluster. This method exploits the fact that label-noise corrupted samples will contain feature-label inconsistencies leading to cluster-label inconsistencies.

Moreover, Isolation Forest (IF) and Local Outlier Factor (LOF) are also used to detect outliers from feature distributions. Both of the above approaches are again used on a per-class basis, training different models on each class subset to better learn class-conditional outlier behavior.

For both IF and LOF, the contamination parameter that denotes the anticipated fraction of outliers is set directly to correspond to the known ratio of injected noise. This environment, though unrealistic for real-world use, is employed here only to facilitate reasonable baseline comparisons in idealized situations.

In addition to testing the methods in isolation, the baseline phase also tests numerous combinations of detection signals. These combinations take logical AND intersections, where an example needs to be flagged by multiple detectors at once in order to count as noisy. These configurations are experimented with, Ensemble plus K-Means++, Ensemble plus Isolation Forest, Ensemble plus LOF, and the complete intersection Ensemble plus K-Means++ and Isolation Forest. These combinations are attempted to check whether the merging of various different perspectives in an integrated way can improve detection rates by reducing false positives, but at the cost of recall.

All the methods are evaluated with respect to the known noisy instances using precision and recall metrics. Precision gives the proportion of the instances that have been discovered and are indeed noisy, while recall provides the proportion of the instances found successfully by the method as noisy.

While the baseline setting offers valuable insight into the intrinsic behavior of each detector, it is quickly realized that employing default hyperparameters is not the best choice for accomplishing optimal detection performance. The table 4 and 5 represent the mean precision, recall and F1-Score per dataset for iteration 1, and also for each method during iteration 1, respectively.

Table 4 - Mean Precision, Recall and F1-Score per Dataset for Iteration 1

Dataset	Noise Level (%)	Precision	Recall	F1-Score
Balance Scale	10	0.439	0.598	0.442
	20	0.508	0.591	0.506
	30	0.600	0.612	0.578
Breast Cancer	10	0.600	0.513	0.512
	20	0.680	0.566	0.591
	30	0.736	0.589	0.631
Car Evaluation	10	0.482	0.541	0.445
	20	0.552	0.538	0.502
	30	0.636	0.533	0.551
Dermatology	10	0.724	0.814	0.747
	20	0.778	0.815	0.783
	30	0.793	0.824	0.797
Glass	10	0.157	0.409	0.193
	20	0.333	0.534	0.374
	30	0.425	0.604	0.475
Iris	10	0.728	0.841	0.759
	20	0.762	0.758	0.744
	30	0.796	0.800	0.788
Phishing	10	0.520	0.360	0.339
	20	0.586	0.385	0.403
	30	0.630	0.411	0.451
Satimage	10	0.507	0.654	0.529
	20	0.585	0.651	0.591
	30	0.651	0.685	0651
Seeds	10	0.727	0.857	0.763
	20	0.738	0.813	0.761
	30	0.770	0.781	0.766
User Knowledge	10	0.442	0.622	0.449
	20	0.499	0.635	0.521
	30	0.537	0.684	0.578
Vertebral	10	0.527	0.565	0.478
	20	0.579	0.593	0.542
	30	0.582	0.601	0.569
Wine	10	0.728	0.836	0.769
	20	0.802	0.838	0.810
	30	0.805	0.822	0.804

Table 5- Mean Precision, Recall and F1-Score by method for iteration 1

Method	Precision	Recall	F1-Score
Ensemble Learning	0.512	0.948	0.647
Ensemble + K-Means++	0.626	0.829	0.647
Ensemble + K-Means++ + Isolation Forest	0.819	0.487	0.589
Ensemble + Isolation Forest	0.732	0.530	0.604
Ensemble + LOF	0.697	0.484	0.557
K-Means++	0.482	0.855	0.594
Isolation Forest	0.525	0.541	0.533
Local Outlier Factor	0.485	0.499	0.491

4.2.6.2 Iteration two – Isolated Hyperparameter Optimization

The second iteration is thus solely for hyperparameter tuning for individual models only, without including any combinations at this time. This intentional segregation is what makes the true contribution of optimized parameterization possible to be tested independently of the potential impact of interaction of method integration.

For the ensemble learning detector, the decision threshold applied to the ensemble confidence scores is no longer fixed. Instead, a grid search over a range of threshold values between 0.3 and 0.9 is performed, and the threshold that maximizes the F1-score on the noisy indices is selected. This tuning acknowledges that a fixed threshold cannot possibly achieve the optimal precision-recall trade-off across different datasets and noise levels.

Also, in the Isolation Forest case, the parameter contamination is set using grid search over values between 0.01 and 0.20. Contamination is directly influencing the sensitivity of the detector, and optimal values must vary across datasets due to class structure and feature separability variation.

Local Outlier Factor also undergoes an identical contamination optimization. Because of the extreme sensitivity of density-based methods like LOF to contamination assumptions, this parameter is adjusted to support varying dataset characteristics.

Hyperparameter tuning is not performed with K-Means++. This choice is both a consequence of the unsupervised learning character of clustering and of the deterministic character of K-Means++ once initialized using the k-means++ method.

Following hyperparameter tuning, precision and recall metrics are derived for each detector optimized separately. Isolated evaluation here provides a more accurate estimate of the true potential of each model under realistic tuning, free from the confounding influence of shared

detection logic. The table 6 and 7 represent the mean precision, recall and F1-Score per dataset for iteration 2, and also for each method during iteration 2, respectively.

Table 6- Mean Precision, Recall and F1-Score per Dataset for Iteration 2

Dataset	Noise Level (%)	Precision	Recall	F1-Score
Balance Scale	10	0.415	0.633	0.411
	20	0.473	0.614	0.486
	30	0.552	0.629	0.557
Breast Cancer	10	0.483	0.683	0.554
	20	0.580	0.624	0.600
	30	0.618	0.591	0.597
Car Evaluation	10	0.376	0.642	0.450
	20	0.449	0.541	0.478
	30	0.481	0.514	0.476
Dermatology	10	0.672	0.864	0.751
	20	0.741	0.849	0.789
	30	0.830	0.778	0.788
Glass	10	0.148	0.486	0.213
	20	0.334	0.606	0.407
	30	0.436	0.594	0.475
Iris	10	0.641	0.850	0.705
	20	0.741	0.767	0.740
	30	0.821	0.739	0.757
Phishing	10	0.306	0.493	0.356
	20	0.392	0.470	0.414
	30	0.452	0.470	0.448
Satimage	10	0.426	0.684	0.492
	20	0.507	0.693	0.571
	30	0.619	0.660	0.617
Seeds	10	0.683	0.917	0.773
	20	0.734	0.827	0.773
	30	0.793	0.726	0.746
User Knowledge	10	0.326	0.776	0.437
	20	0.459	0.689	0.519
	30	0.558	0.636	0.557
Vertebral	10	0.340	0.653	0.433
	20	0.463	0.585	0.498
	30	0.565	0.551	0.540
Wine	10	0.697	0.984	0.815
	20	0.779	0.875	0.824
	30	0.791	0.745	0.757

Table 7- Mean Precision, Recall and F1-Score by method in iteration 2

Method	Precision	Recall	F1-Score
Ensemble Learning	0.678	0.888	0.755
K-Means++	0.441	0.820	0.548
Isolation Forest	0.548	0.525	0.524
Local Outlier Factor	0.519	0.482	0.485

4.2.6.3 Iteration three – Integrated Detection through Optimized Combinations

The third iteration remains to explore more sophisticated integration strategies. In this case, the optimized detectors are incorporated into multi-stage hybrid schemes to take advantage of supervised and unsupervised detection paradigms' complementary strengths.

The optimum parameters for ensemble confidence thresholds and contamination levels are reused wholesale for consistency and to enable a fair comparative evaluation. Different integration architectures are then constructed.

In the Ensemble + K-Means++ combination, the instance is marked as noisy only if both detectors raise an alarm on their own, thus conservatively adopting an intersection approach. The same is adopted for the Ensemble + Isolation Forest and Ensemble + LOF combinations. A voting mechanism approach is adopted in the new Ensemble + K-Means++ + Isolation Forest combination and that is more liberal. Here, an example is reported if at least two out of the three detectors have concurred on its noisiness, providing a trade-off between overconservative and over permissive decision logic.

Finally, a Two-Stage architecture is looked at, drawing inspiration from multi-level noise filtering architectures in the literature. In this system, the previously tuned ensemble model is first employed to report high-confidence noise candidates. They are then submitted through secondary confirmation through either K-Means++ or Isolation Forest. The sample is not tagged until confirmed by a secondary detector. The tiered design is such that the ensemble model's high recall is addressed while precision is improved through the confirmation filters.

As in all previous versions, metrics for evaluation are computed for every detection method at all noise levels. The consistent application of the same noise injection protocols, parameter grids, and evaluation metrics ensures that the comparison during the experiments is robust and easily interpretable.

This incremental experimental configuration therefore enables a comprehensive assessment of the individual and combined efficacy of the new approaches and provides empirical evidence to support further analysis and conclusions. The table 8 and 9 represent the mean precision, recall and F1-Score per dataset for iteration 3, and also for each method during iteration 3, respectively.

Table 8- Mean Precision, Recall and F1-Score per Dataset for Iteration 3

Dataset	Noise Level (%)	Precision	Recall	F1-Score
Balance Scale	10	0.473	0.600	0.459
	20	0.547	0.585	0.524
	30	0.649	0.603	0.592
Breast Cancer	10	0.710	0.682	0.667
	20	0.768	0.640	0.670
	30	0.749	0.612	0.643
Car Evaluation	10	0.526	0.537	0.493
	20	0.562	0.462	0.482
	30	0.570	0.477	0.493
Dermatology	10	0.813	0.881	0.835
	20	0.836	0.853	0.837
	30	0.900	0.799	0.829
Glass	10	0.201	0.505	0.267
	20	0.398	0.627	0.462
	30	0.503	0.598	0.512
Iris	10	0.755	0.842	0.780
	20	0.867	0.773	0.807
	30	0.899	0.747	0.798
Phishing	10	0.572	0.444	0.442
	20	0.645	0.429	0.468
	30	0.649	0.432	0.482
Satimage	10	0.577	0.720	0.599
	20	0.664	0.727	0.667
	30	0.752	0.693	0.691
Seeds	10	0.749	0.922	0.813
	20	0.789	0.820	0.795
	30	0.846	0.736	0.773
User Knowledge	10	0.498	0.782	0.567
	20	0.613	0.689	0.608
	30	0.685	0.630	0.618
Vertebral	10	0.523	0.636	0.540
	20	0.655	0.576	0.572
	30	0.758	0.541	0.599
Wine	10	0.772	0.983	0.862
	20	0.878	0.882	0.872
	30	0.888	0.767	0.802

Table 9- Mean Precision, Recall and F1-Score by method for iteration 3

Method	Precision	Recall	F1-Score
Two-Staged	0.729	0.818	0.757
Ensemble Learning	0.678	0.888	0.755
Ensemble + K-Means++	0.728	0.766	0.729
Ensemble + K-Means++ + Isolation Forest	0.648	0.833	0.719
Ensemble + Isolation Forest	0.824	0.496	0.604
Ensemble + LOF	0.816	0.447	0.559
K-Means++	0.441	0.820	0.548
Isolation Forest	0.548	0.525	0.524
Local Outlier Factor	0.519	0.482	0.485

4.2.7 Noise Mitigation with Two-Staged Filtering

Following the step-by-step advancement of the noise detection techniques in the above steps, this final step presents the complete mitigation process. The purpose of this step is no longer just to discover noisy instances, but to use the installed noise detection system as a preprocessing filter, removing the noisy samples identified from the training set before retraining the supervised models. This process allows for testing, whether actively removing noisy samples from training by active filtering, can improve supervised learning models' generalization performance under noisy labels.

The mitigation phase is based entirely on the iteration three architecture, leveraging its most powerful and efficient setup. For example, the Two-Stage detection approach is used as the primary filtering method to identify noise using a cascade of supervised and unsupervised approaches. The Two-Stage design features high recall and precision in labelling potential mislabeled data and minimizes deletions of well labelled data unnecessarily.

The experiment begins with injecting systematic label noise in controlled amounts into the training set, simulating realistic class noise scenarios at ten, twenty and thirty percent noise levels. The noise injection technique is formulated to instruct class-balanced perturbations such that an equal ratio of instances is flipped per class. This stratified approach preserves class distribution representativeness within experimental testing and prevents class imbalance artifacts from distorting the noise effects.

Prior to the filtering itself being applicable, calibration of the Two-Stage detection model on each dataset and noise level is necessary. This tuning process involves discovery of two hyperparameters that regulate the activities of both stages of the system. The first one is the

ensemble threshold, i.e., the minimum confidence level that instances are labelled as suspicious by the ensemble model in the first phase. The second one controls the contamination rate used by the Isolation Forest model in phase two and acts as a control for its sensitivity when identifying anomalous patterns in the feature space per class.

The ensemble threshold tuning is done on a grid search basis, iterating through values from 0.3 to 0.9 in increments and observing the resulting precision, recall, and F1-score on noise detection against the artificially added noise as ground truth. This helps us decide on the threshold that gives the best trade-off between correctly detecting noisy instances without too many false positives. An identical grid search process is performed to tune the contamination rate of the Isolation Forest. This unsupervised detector is used individually within every class label, keeping class-conditional patterns intact at anomaly detection, and its contamination parameter is chosen to optimize its independent F1-score on noise detection.

Once the optimal parameters are determined for a particular dataset and noise level, the overall noise detection process is executed. During the first phase, ensemble confidence scores are calculated as the mean predicted probabilities of the three calibrated classifiers, SVM, GNB and RF, replicating the mechanism set in earlier iteration one. Samples with confidence scores less than the optimized threshold are labelled as primary noise candidates.

In the second stage, more anomaly detection models are used to narrow down the initial candidate set. Isolation Forest and Local Outlier Factor models using previously optimized contamination rates are used within each class in isolation to identify within-class inconsistencies that could be an indication of label corruption. In addition, K-Means++ clustering is utilized to assess class-cluster consistency, by mapping unsupervised clusters onto recognized class labels via majority assignment, and identifying cluster assignment vs. noisy label disputes.

The final Two-Stage filtering approach decision rule operates by accumulating these few detection signals. I.e., a sample reported by the ensemble in Phase 1 is retained as noisy only if one of the Phase 2 algorithms, Isolation Forest or K-Means++, individually confirms the suspicion. The cascaded nature of this reasoning filters out only examples exhibiting both supervised uncertainty and unsupervised inconsistency, exchanging sensitivity and specificity in noise filtering.

After the filtering process, the filtered training set is used to retrain the new Random Forest classifier. This model is then tested on the original, noise-free test set that remains constant across all experimental conditions for comparison. This post-filtering performance is systematically compared with baseline performance obtained from training on noisy data without filtering, thereby allowing the quantification of potential benefits of the proposed mitigation strategy.

The actions of these mitigation experiments are the best threshold and contamination hyperparameters selected for a particular setup, the number of instances remaining after filtering, and the retraining test accuracy achieved after retraining. These are summed up systematically across all datasets and levels of noise to provide an overall assessment of the applied benefit of doing noise mitigation using the proposed Two-Stage filtering framework.

5 Results and Discussion

5.1 Evaluation of Noise Detection Performance

The effectiveness of the noise-detection framework was assessed across the twelve datasets detailed in the previous section. Each dataset was subject to three level of synthetic label corruption and over three iterations, beginning with the baseline calibrated ensemble, progressing to the systematic hyperparameter optimization and ending with the two-staged hybrid approach, and it is possible to conclude that the last one showed a clear and consistent enhancement in overall F1-Scores.

The hyperparameter tunning itself conceded an improvement, raising the ensemble learning F1-Score mean from 0.647 up to 0.75. However, the standalone unsupervised detectors continued with low values of mean F1-score, bellow 0.600, even after the tunning step. Integrating these methods with the two-stage approach showed an uplift, reaching the peak mean F1-Score of 0.757.

Table 10 summarizes the mean F1-Score values for each method used with different noise levels, across the three iterations implemented, illustrating the gains obtained through tuning and the use of the hybrid approach.

Table 10- Global mean performance of noise detection across all datasets and noises

Method	Iteration 1 F1-Score	Iteration 2 F1-Score	Iteration 3 F1-Score
Ensemble Learning	0.647	0.755	0.754
Ensemble Learning + K-Means ++	0.693	---	0.729
Ensemble Learning + Isolation Forest	0.604	---	0.604
Ensemble Learning + K-Means ++ + Isolation Forest	0.589	---	0.719
Two-Staged	---	---	0.757
K-Means ++	0.594	0.548	0.548
Isolation Forest	0.533	0.524	0.524
Local Outlier Factor	0.491	0.485	0.485

5.2 Performance Analysis per Group

To smooth the analysis of the results, the twelve datasets were separated in three behaviorally distinct groups, each one composed by four datasets, so that similarities in feature separability, class balance and response to noise become apparent.

The table 11 describes the organization of each group of datasets. The group A is formed by the Iris, Wine, Dermatology and Seeds datasets. These have clear boundaries and low overlap, accommodating consistently strong performance. The Group B has the Balance Scale, Breast Cancer, Car Evaluation and Vertebral Column, since the classes are moderately overlapping and have features with medium complexity. The last one, the group C, has Glass, Phishing, Satimage and User Knowledge, and exhibit high class overlap, class imbalance and high dimensional structures that complicate noise detection.

Table 11- Datasets separation per group

Group	Dataset	Characteristics
A	Iris, Wine, Dermatology, Seeds	High-Separability, well structed
B	Balance Scale, Breast Cancer, Car Evaluation, Vertebral Column	Moderate-Separability, mid-sized complexity
C	Glass, Phishing, Satimage, User Knowledge	Low-Separability, challenging

5.2.1 Group A

The Group A is portrayed by well separated classes and low over position in the distribution characteristics. For that reason, even on the first iteration it is possible to see a high mean F1-score value. The table 12 resumes the mean of the metrics of precision, recall and F1-score for each level noise and iteration.

Table 12- Group A's mean precision, recall and F1-Scores per iteration and noise levels

Noise Levels	Metrics	Iteration 1	Iteration 2	Iteration3
10%	Precision	0.727	0.673	0.772
	Recall	0.837	0.904	0.907
	F1-Score	0.759	0.705	0.780
20%	Precision	0.770	0.749	0.843
	Recall	0.806	0.830	0.832
	F1-Score	0.783	0.789	0.837
30%	Precision	0.791	0.809	0.883
	Recall	0.807	0.747	0.762
	F1-Score	0.797	0.788	0.829

At ten percent noise the iteration 1 achieves a solid F1-Score of 0.759, reflecting balanced precision, 0.727, and recall, 0.837. The introduction per dataset of the hyperparameters tuning, in iteration 2, shifts this balance. Precision falls to 0.673 as the decision threshold tightens against false positives, while recall climbs up to 0.904. Yet the real effect is on the F1-score that drops to 0.705. To contrast this situation, during iteration 3, two stage recovers both precision, that rises to 0.772, and recall, that improves to 0.907, lifting the F1-Score to 0.780. This showed that the hybrid confirmation can sharpen selectivity without discarding sensitivity.

Under twenty percent of noise, the baseline detectors' F1-Score of 0.783 already says a lot regarding the quality of the group. Tuning alone contributes with a slight increase to 0.789 by raising recall to 0.830, but at the expense of a minimal reduction in precision. With the two-stage architecture, precision is raised up to 0.843 with recall still at 0.832, with a substantial F1-Score of 0.837 (+0.054 over baseline). This advantage guarantees that, under moderate noise conditions, ensembling uncertainty with consistency in feature space is the largest performance advantage.

With the level of noise increased to thirty percent, baseline precision, 0.791, and recall, 0.807, sum up to the highest early F1-Score of 0.797. Hyperparameter tuning in Iteration 2, gives, once again, some recall improvements, to 0.747, for greater precision, 0.809, resulting in a slightly lower F1-Score of 0.788. The final hybrid phase shifts the emphasis significantly in favor of precision, 0.883, with recall recovering to some extent to 0.762. The resulting F1-Score of 0.829 remains a high value, even if not as good as the moderate noise optimum. This move illustrates the well-known precision/recall trade-off in high noise environments. The two-stage filter

sacrifices a fraction of sensitivity, in order to eliminate false positives at the cost of discarding a minor fraction of noisy true cases.

In summary, Group A's clear class separation facilitates uniformly good baseline detection. Isolated hyperparameter tuning achieves small gains or losses, while two-stage hybrid achieves the biggest F1-Score gains, particularly at intermediate noise levels, by mixing calibrated ensembles' sensitivity with unsupervised confirmation's specificity.

5.2.2 Group B

The second group presents moderate class overlap and a mix of categorical and numerical features, which can turn noise detection more challenging when compared to the previous group A. The table 13 shows values for mean precision, recall and F1-Scores across all iterations for the different noise levels

Table 13- Group B ’s mean precision, recall and F1-Scores per iteration and noise levels

Noise Levels	Metrics	Iteration 1	Iteration 2	Iteration3
10%	Precision	0.511	0.408	0.562
	Recall	0.574	0.659	0.635
	F1-Score	0.442	0.411	0.459
20%	Precision	0.581	0.494	0.639
	Recall	0.588	0.611	0.598
	F1-Score	0.506	0.486	0.524
30%	Precision	0.641	0.567	0.696
	Recall	0.604	0.589	0.585
	F1-Score	0.578	0.557	0.592

At ten percent noise, baseline detector works reasonably well, with a F1-Score of 0.442, low precision, 0.511, and recall, 0.574, in balance. Hyperparameter tuning in Iteration 2 enhances recall to 0.659 by lowering the ensemble confidence threshold and contamination levels, but doing so, ends up lowering the precision, 0.408, and there is a net loss of the F1-Score to 0.411. The two-stage cascade of Iteration 3 both captures aspects of performance, precision is improved to 0.562 and recall remains high at 0.635, raising theF1-Score to 0.459 and demonstrating that unsupervised validation can reverse the too many false positives induced by tuning alone.

At the twenty percent noise level, the baseline setting achieves F1-score of 0.506. Iteration 2's parameter adjustment again increases recall to 0.611, but also pushes down precision to 0.494 so that the F_1-Score is lowered slightly to 0.486. Iteration 3's hybrid detection, however, achieves precision of 0.639 and recall 0.598, giving an F1-Score of 0.52, a 0.018 increase over

the baseline. This result demonstrates the two-stage model's ability to recover selectivity without sacrificing sensitivity even in the presence of increased noise.

At thirty percent noise, baseline F1-Score is 0.578, which means higher noise prevalence paradoxically improves the detector's performance in recognizing mislabeled patterns in these sets. Iteration 2 also compromises precision, 0.567, and recall, 0.589, with F1-Score of 0.557. Iteration 3 benefits from both, precision improves to 0.696 and recall continues at 0.585, with an F1-Score of 0.592. Although this is below baseline peak, it is nonetheless superior to the stand-alone, it demonstrates the hybrid approach's steady advantage in reducing false positives and false negatives.

Briefly, Group B's less diversified feature spaces by definition complicate noise detection. Hyperparameter tuning only leads to over-compensation that boosts recall at the expense of precision. The two-stage cascade of Iteration 3 only improves consistently over the baseline F1-Score on all noise levels by harnessing ensemble uncertainty and unsupervised validation in feature space.

5.2.3 Group C

The last group poses the greatest challenge due to its pronounced class overlap, heterogenous feature scales and complex decision boundaries. The table 14 summarizes the mean values of precision, recall and F1-Scores across all iterations and noise levels.

Table 14- Group C ’s mean precision, recall and F1-Scores per iteration and noise levels

Noise Levels	Metrics	Iteration 1	Iteration 2	Iteration3
10%	Precision	0.373	0.260	0.424
	Recall	0.464	0.585	0.577
	F1-Score	0.193	0.213	0.267
20%	Precision	0.473	0.395	0.552
	Recall	0.518	0.588	0.582
	F1-Score	0.374	0.407	0.468
30%	Precision	0.531	0.482	0.612
	Recall	0.566	0.567	0.553
	F1-Score	0.475	0.475	0.512

At ten percent noise, the baseline detector gives very poor discrimination with a F1-Score of 0.193, whereas recall is moderate, 0.464, precision remains low, 0.373, which indicates fewer than four in ten labeled as noisy are indeed noisy. Including hyperparameter tuning in Iteration 2 shifts the operating point, recall rises to 0.585 but at the steep cost of precision, 0.260, and the overall F1-Score rises scarcely at all to 0.213, displaying an obsessive false-positive rate. The two-stage cascade of Iteration 3 delivers a more even result as precision rises to 0.424 and recall remains high at 0.577, taking the F1-Score to 0.267. This thirty-eight percent relative gain above

baseline F1-Score, shows that unsupervised confirmation represses much of the over-flagging resulting from tuning alone.

As noise is increased to twenty percent, baseline properly flags nearly half of noisy labels, recall 0.518, with corresponding precision of 0.473 and F1-Score of 0.374. Iteration 2 tuning brings recall up to 0.588 but reduces precision to 0.395, improving only slightly to F1-Score, 0.407. In contrast, Iteration 3's fusion detection greatly enhances selectivity, accuracy improves to 0.552 while remembering 0.582, driving the F1-Score to 0.468, a persistent 0.094 point increase over the untuned baseline.

Using the largest thirty percent noise level, baseline performance with F1-Score of 0.475, records the detector's difficulty in distinguishing true structure from the pervasive label corruption. Hyperparameter adjustment in Iteration 2 does not result in any F1-Score improvement, it remains at 0.475, while there are minor improvements in precision. In contrast, Iteration 3's two-stage structure increases precision to 0.612 while recall is kept at 0.553, with an F1-Score of 0.512. Although this is less than that for less noisy groups, it again points out the value of the hybrid method: even with high levels of noise, coupled supervised–unsupervised filtering recovers more than eight F_1 points from the baseline.

Overall, high feature heterogeneity and overlapping classes of Group C severely degrade out-of-the-box noise detection. Hyperparameter tuning alone oscillates between under- and over-sensitivity, whereas the combined two-stage cascade maximizes systematically the precision vs. recall balance at all noise levels. These results all validate that systematically label noise can be reliably identified only by combining ensemble uncertainty with unsupervised confirmation in feature space.

5.3 Evaluation of Mitigation Effectiveness

After evaluating the performance of noise detection in each of the iterations of the proposed detection framework, the final task of this experimental study was to evaluate if the process of noise mitigation can actually contribute to improvement in the performance of the downstream supervised learning task.

The mitigation process, executed by the proposed iteration three Two-Stage noise detection approach, is to take the optimized noise detection models, identify suspected noisy instances, filter them from the training data, and retrain a downstream classifier on the filtered data. As noted above, the Random Forest classifier was employed as the common evaluation model for all experiments to allow comparability between baseline and post-mitigation performances.

5.3.1 Global Evaluation of Accuracy Gains During Mitigation Step

To comparatively assess the mitigation impact across the different datasets and noise levels, it was estimated the accuracy gain, such as the difference in accuracy attained after mitigation and the baseline accuracy attained during training on the noisily labelled data itself.

Table 15- Summary of aggregated accuracy gain per dataset and noise level

Group	Dataset	10% Noise	20% Noise	30% Noise
A	Iris	-4.44%	+2.22%	+8.89%
	Wine	+1.85%	+3.70%	+1.85%
	Dermatology	-0.91%	-0.91%	-3.64%
	Seeds	+4.76%	+3.17%	+11.11%
B	Balance Scale	+4.79%	+4,26%	+6.91%
	Breast Cancer	-1.17%	0.00%	+4.68%
	Car Evaluation	+2.12%	+5.20%	+0.50%
	Vertebral	+4.30%	+2.15%	+3.23%
C	Glass	-7.69%	+1.54%	-7.69%
	Phishing	+0.91%	3.87%	+1.42%
	Satimage	-1.55%	1.35%	-1.66%
	User Knowledge	-3.31%	+0.83%	+0.83%

As has been observed in table 15, the mitigation strategy also shows encouraging accuracy gains in most of the datasets, particularly at the higher noise rates of twenty percent and thirty percent. This confirms that the proposed technique was able to successfully identify and remove noisy mislabeled instances and enhance model performance in heavily label noise-contaminated cases.

However, certain datasets such as Glass, Dermatology, Satimage, and User Knowledge to some degree exhibit little or even a negative improvement over certain levels of noise, these kinds of fluctuations are likely to be related to the inherent complexity, skewness and small data size of

these specific datasets that make it challenging to determine noise and filtering without suffering collateral loss of informative examples.

5.3.2 Visual Analysis

The figures bellow provide two graphical representations of accuracy gain distribution across datasets and noise levels.

The figure 2 presents the accuracy gain pattern per dataset as noise levels increase

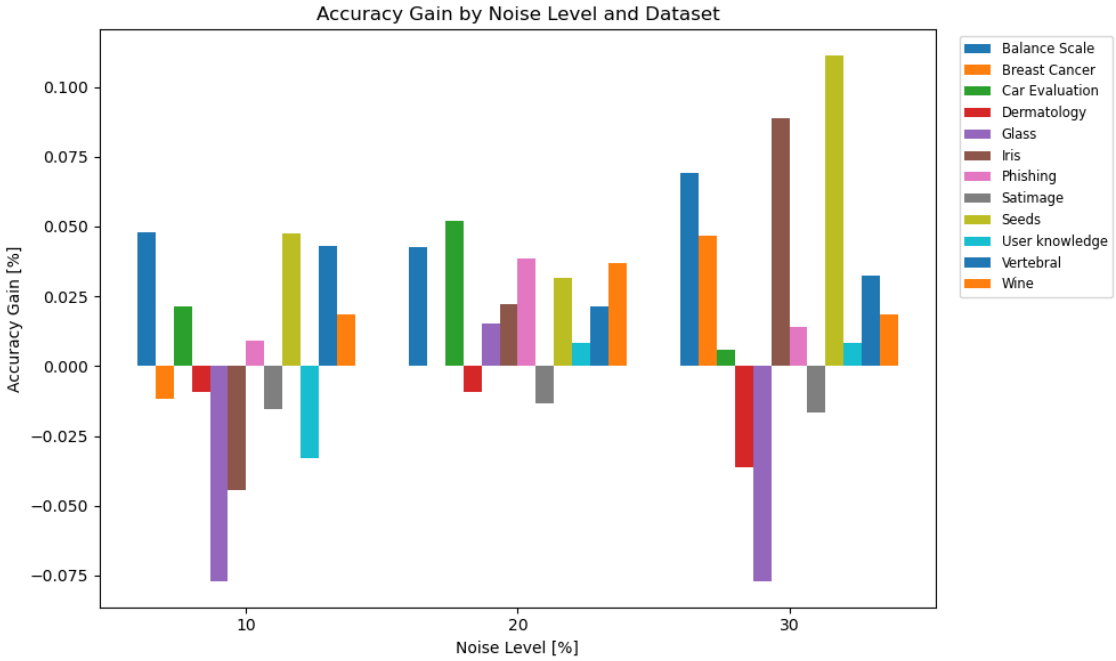

Figure 2- Accuracy Gain vs Noise Level Histogram

The histogram shows that data sets like Seeds, Iris, and Balance Scale show consistent improvement in accuracy with higher levels of noise, indicating the robustness of the mitigation algorithm in these data sets. In contrast, data sets like Dermatology and Glass show erratic

behavior, suggesting that there has been a time when over-filtering has removed some useful clean instances along with noisy ones.

The second plot, figure 3, is a heatmap of accuracy gains:

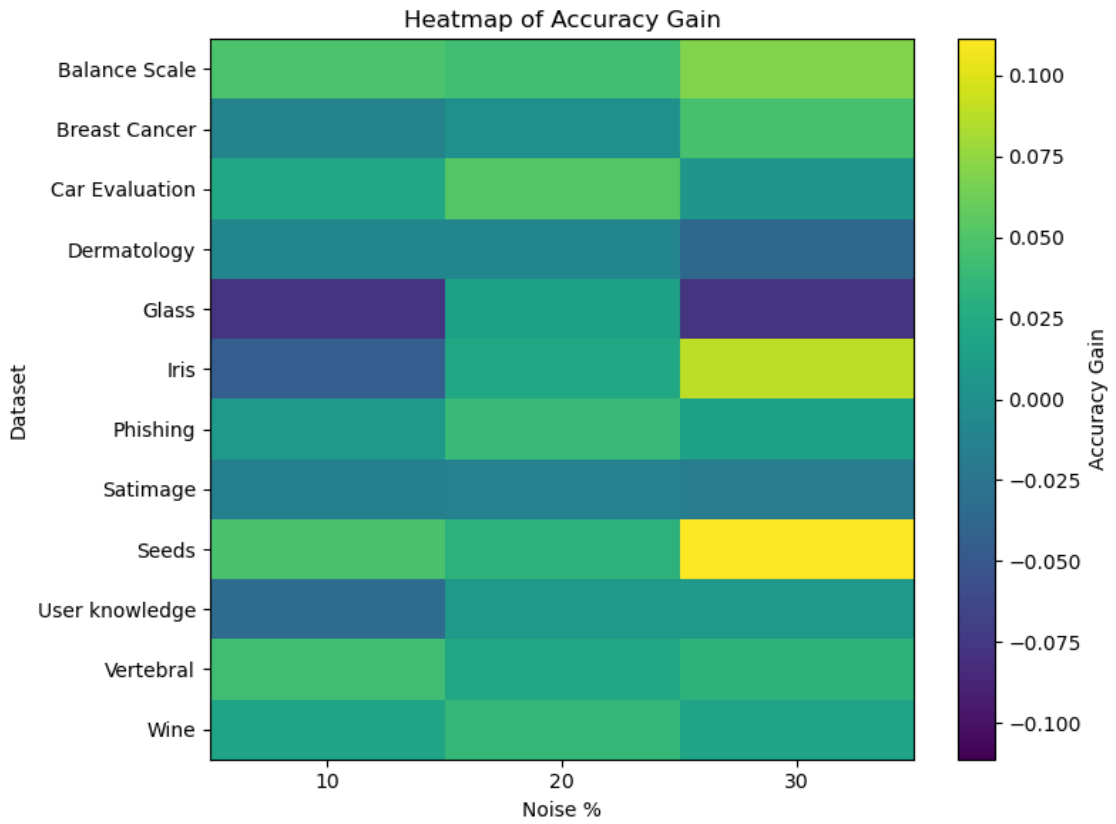

Figure 3- Heatmap of Accuracy Gains

The heatmap is particularly good at picking out the datasets that benefited most from the mitigation strategy, Seeds to +11 percent gain at thirty percent noise, while picking out tougher cases where the noise filtering had minimal effect or actually decreased performance.

5.3.3 Global Interpretation

In general the results suggests that the proposed two-stage noise filtering method is quite effective, particularly in the event of moderately distributed classes, satisfactory sample size, and relatively separated classes.

The larger the level of noise, the more the effect of filtering, such as in several datasets where improvement was greater with higher levels of noise contamination.

Datasets with excessive class overlap, imbalance, or limited training instances, such as Glass and Dermatology, remain challenging and may require more specific preprocessing, or more

judicious filtering thresholds or more hybrid methods involving detection with confidence-based relabeling instead of flat instance removal.

These findings validate the hypothesis that systematic label noise could be effectively addressed by tight detection and filtering mechanisms and emphasize the importance of hyperparameter tuning at each step of the mitigation pipeline to acquire maximum robustness.

5.4 Final Discussion

The comprehensive experimental framework developed throughout the duration of this work enabled a thorough investigation of supervised learning robustness to label noise, and in particular systematic in nature. The multi-phase experimental design, with incremental algorithmic revisions, iteration one to three, systematic hyperparameter tuning, and subsequent mitigation analysis, allows for several high-level observations to be made.

Firstly, the incremental complexity of the noise detection pipeline helped achieve more and more robust detection results. The initial confidence-based ensemble detector, iteration one, already provided a solid foundation, particularly due to the diversity generated by the combination of probabilistic predictions from various classifiers. However, its lack of dataset-specific specialization limited it from generalizing well to diverse datasets and noise levels.

The second revision, iteration two, by explicit hyperparameter tuning, demonstrated that diligent ensemble threshold and anomaly contamination rate tuning for Isolation Forest and LOF could dramatically enhance detection recall and precision. This only goes to highlight that noise detection is acutely sensitive to dataset-specific structures, class distributions, and underlying feature spaces.

The final revision, iteration three, consolidated these enhancements into a hybrid two-stage detection system, cascading supervised and unsupervised detectors. The two-stage method leveraged the high confidence of ensemble predictions to define an initial candidate set, which was then purified by feature consistency analysis, K-Means++, Isolation Forest, LOF. This process consistently yielded higher F1-scores, indicating better balance between recall and precision. Remarkably, this framework also demonstrated stability at high noise levels, thirty percent, where most traditional methods begin deteriorating severely.

Apart from detection quality, the mitigation analysis also revealed strong evidence of downstream practical utility. By filtering out examples detected as noisy, subsequent supervised model training on the cleaned datasets frequently led to improved classification accuracies. Yet, the magnitude of the improvements was not uniform across datasets, with some of them, for example, Seeds, Balance Scale, Iris and Wine, receiving clear gains while others, Glass, Dermatology and Satimage, sometimes negatively suffered from inadvertent over-filtering when class overlap or small sample sizes were present.

Together, the results validate the primary hypothesis of this dissertation: that systematic label noise can be effectively detected and mitigated under heterogeneous ensemble methodologies, provided that sufficient model diversity, per-dataset calibration, and multi-phase verification mechanisms are employed. Furthermore, this work demonstrates that robust noise handling is

not merely a supplementary preprocessing step but has a direct impact on the downstream generalization performance of supervised learning models.

These findings suggest several avenues for further development, such as the application of adaptive thresholding techniques, dynamic ensemble member weighting, and further exploration of semi-supervised or active learning strategies to enhance detection boundaries. Nevertheless, the framework developed here provides a solid foundation for establishing supervised learning robustness to real-world noisy data conditions, with direct applicability to practical domains such as healthcare, finance, and social sciences where annotation quality tends to be less than perfect.

6 Conclusions and Future Directions

6.1 Conclusions

This study has demonstrated that a two-step hybrid pipeline, combining a calibrated ensemble of Support Vector Machine, Random Forest and Naïve Bayes classifiers with k-Means++ clustering and class-conditional anomaly detectors, improves detection and mitigation of systematic label noise on a dozen heterogeneous benchmark datasets. In a baseline configuration, the calibrated ensemble alone achieved an average F1-Score of 0.647 under exposure to ten, twenty and thirty percent synthetic noise. Through systematic hyperparameter tuning of the ensemble confidence threshold and anomaly contamination rates, that F1-Score increased to 0.755. Lastly, by cascading supervised uncertainty estimation with unsupervised consistency checking in a two-stage framework, this approach achieved an overall F1-Score of 0.757, outperforming every single-stage approach and reducing both false positives and false negatives.

Datasets with well-separated classes and low overlap, such as Iris, Wine, Dermatology and Seeds, also exhibited the best baseline noise-detection performance with a F1-Score bigger than 0.75, and still benefited from the two-stage pipeline by up to 0.054. On feature domains of medium overlap and mixed feature types, like Balance Scale, Breast Cancer, Car Evaluation and Vertebral Column, where decision boundaries are less clearly defined, the hybrid approach still produced clear F1 gains of 0.018 to 0.066. Most obviously, even highly overlapping, high-dimensional datasets, such as Glass, Phishing, Satimage and User Knowledge, which suffered the worst out-of-the-box detection, with a F1-Score smaller than 0.40, had relative improvements of up to 0.094 when ensemble predictions were ratified through unsupervised anomaly and clustering cues. These findings confirm that feature complexity and class overlap strongly detract from noise detection, yet a carefully tuned two-stage pipeline recovers performance under all evaluated conditions.

When the two-stage filter was employed as a preprocessing step to remove flagged samples before retraining a Random Forest classifier, predictive accuracy gains of up to 11.11% were

observed, in the Seeds dataset with the noise level at thirty percent, with more moderate improvements of two to five percent on the majority of datasets at the higher noise levels. This confirms that effective noise mitigation not only cleanses corrupted labels but also results in better downstream generalization. However, in regions of extensive overlap or small sample sizes, Glass, Dermatology, Satimage, User Knowledge, vigorous filtering sometimes removed informative examples, leading to modest declines in accuracy. This points to the need for tailoring confidence thresholds and contamination parameters to the peculiarities of each data set.

In conclusion, the evidence shows that the proposed hybrid approach outperforms existing baseline approaches, is robust against a wide range of dataset complexities, and largely enhances predictive accuracy under systematic label corruption.

6.2 Limitations and Future Work

In spite of these encouraging results, this study also uncovered some limitations that can be explored further. In datasets that show severe class imbalance, extremely high feature dimensionality or strong boundary overlap, for example Glass, Dermatology and Satimage, the two-stage filter at times deleted informative examples as well as noisy examples, resulting in suboptimal or even negative improvements in subsequent accuracy. This phenomenon evidences the intrinsic trade-off between sensitivity and specificity when using deletion-based mitigation.

Follow-up work should therefore explore more subtle strategies for handling uncertain cases. One possibility is confidence-association relabeling instead of removal, thereby keeping potentially useful data while still penalizing low-confidence labels. Adaptive ensemble techniques that update classifier weights flexibly based on per-class calibration performance could further improve robustness in imbalanced domains. Enlarging the pipeline to accommodate more realistic noise models, such as feature-dependent or adversarial labeling errors, would enhance its applicability to real-world problems further.

Finally, extending this framework to unstructured data modalities, like text, images, time series, and to streaming environments where labels arrive sequentially, would make it even more applicable to modern machine learning tasks. Investigation of semi-supervised or active learning extensions, where human information can be selectively queried to resolve doubtful cases, offers the promise of trading off automation with intelligent control. Together, these developments will pave the way towards truly noise-resilient prediction machines suitable for high-stakes domains such as healthcare, finance and social sciences.

Bibliography

[1] B. Han *et al.*, “A Survey of Label-noise Representation Learning: Past, Present and Future,” *A SURVEY OF LABEL-NOISE REPRESENTATION LEARNING*, vol. XX, Nov. 2020, Accessed: Jun. 29, 2025. [Online]. Available: https://arxiv.org/abs/2011.04406v2

[2] A. Jobin, M. Ienca, and E. Vayena, “Artificial Intelligence: the global landscape of ethics guidelines,” *Nat Mach Intell*, vol. 1, no. 9, pp. 389–399, Jun. 2019, doi: 10.1038/s42256-019-0088-2.

[3] N. Kluge Corrêa, J. M. Mönig, J. Maria, M. M.-B. De, and L. Hinz, “Catalog of General Ethical Requirements for AI Certification,” Aug. 2024, Accessed: Jun. 29, 2025. [Online]. Available: https://arxiv.org/abs/2408.12289v2

[4] B. Frénay and M. Verleysen, “Classification in the presence of label noise: A survey,” *IEEE Trans Neural Netw Learn Syst*, vol. 25, no. 5, pp. 845–869, 2014, doi: 10.1109/TNNLS.2013.2292894.

[5] H. Song, M. Kim, D. Park, Y. Shin, and J. G. Lee, “Learning from Noisy Labels with Deep Neural Networks: A Survey,” *IEEE Trans Neural Netw Learn Syst*, vol. 34, no. 11, pp. 22–25, Jul. 2020, doi: 10.1109/TNNLS.2022.3152527.

[6] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” *Procedia Comput Sci*, vol. 181, pp. 526–534, 2021, doi: 10.1016/J.PROCS.2021.01.199.

[7] “Grammarly: Free AI Writing Assistance.” Accessed: Jan. 03, 2025. [Online]. Available: https://www.grammarly.com/

[8] K. Nikolaidis, T. Plagemann, S. Kristiansen, V. Goebel, and M. Kankanhalli, “Using Under-trained Deep Ensembles to Learn Under Extreme Label Noise,” Sep. 2020, Accessed: Jun. 29, 2025. [Online]. Available: https://arxiv.org/abs/2009.11128v2

[9] P. Filzmoser, R. Maronna, and M. Werner, “Outlier identification in high dimensions,” *Comput Stat Data Anal*, vol. 52, no. 3, pp. 1694–1711, Jan. 2008, doi: 10.1016/J.CSDA.2007.05.018.

[10] B. E. and F. A., “Identifying mislabeled training data,” *Journal of Artificial Intelligence Research*, Jul. 1999, doi: 10.5555/3013545.3013548.

[11] F. T. Liu, K. M. Ting, and Z. H. Zhou, “Isolation forest,” *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 413–422, 2008, doi: 10.1109/ICDM.2008.17.

[12] C. G. Northcutt, L. Jiang, and I. L. Chuang, “Confident Learning,” *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, Apr. 2021, doi: 10.1613/JAIR.1.12125.

[13] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, “Deep Learning is Robust to Massive Label Noise,” Feb. 2018, Accessed: Jun. 21, 2025. [Online]. Available: http://arxiv.org/abs/1705.10694

[14] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, “Credit card fraud detection: A realistic modeling and a novel learning strategy,” *IEEE Trans Neural Netw Learn Syst*, vol. 29, no. 8, pp. 3784–3797, Aug. 2018, doi: 10.1109/TNNLS.2017.2736643.

[15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013, doi: 10.1177/0278364913491297.

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning (still) requires rethinking generalization,” *Commun ACM*, vol. 64, no. 3, pp. 107–115, Feb. 2021, doi: 10.1145/3446776.

[17] D. Arplt *et al.*, “A Closer Look at Memorization in Deep Networks,” *34th International Conference on Machine Learning, ICML 2017*, vol. 1, pp. 350–359, Jun. 2017, Accessed: Jun. 21, 2025. [Online]. Available: https://arxiv.org/abs/1706.05394v2

[18] G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu, “Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach,” Mar. 2017, Accessed: Jun. 21, 2025. [Online]. Available: http://arxiv.org/abs/1609.03683

[19] J. Goldberger and E. Ben-Reuven, “Training deep neural-networks using a noise adaptation layer,” Feb. 06, 2017.

[20] J. W. Sun, F. Y. Zhao, C. J. Wang, and S. F. Chen, “Identifying and correcting mislabeled training instances,” *Proceedings of Future Generation Communication and Networking, FGCN 2007*, vol. 1, pp. 244–250, 2007, doi: 10.1109/FGCN.2007.146.

[21] A. Vahdat, “Toward Robustness against Label Noise in Training Deep Discriminative Neural Networks”.

[22] Z. H. Zhou, “Ensemble methods: Foundations and algorithms,” *Ensemble Methods: Foundations and Algorithms*, pp. 1–218, Jan. 2012, doi: 10.1201/B12207.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” 2017, doi: 10.5555/3305381.3305518.

[24] R. Hasan and C. H. H. Chu, “A Heterogeneous Ensemble Method for Handling Class Noise in Supervised Machine Learning,” *Proceedings of the ACM Symposium on Applied Computing*, pp. 902–909, Apr. 2024, doi: 10.1145/3605098.3635936.

[25] M. M. Breuniq, H. P. Kriegel, R. T. Ng, and J. Sander, “LOF,” *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93–104, May 2000, doi: 10.1145/335191.335388.

[26] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, “Deep Learning for Anomaly Detection: A Review,” *ACM Comput Surv*, vol. 54, no. 2, Mar. 2021, doi: 10.1145/3439950/SUPPL_FILE/PANG.ZIP.

[27] J. Li, R. Socher, and S. C. H. Hoi, “DivideMix: Learning with Noisy Labels as Semi-supervised Learning,” *8th International Conference on Learning Representations, ICLR 2020*, Feb. 2020, Accessed: Jun. 21, 2025. [Online]. Available: https://arxiv.org/abs/2002.07394v1

[28] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, Jul. 2021, doi: 10.1145/3457607.

[29] C. G. Northcutt, A. Athalye, and J. Mueller, “Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks,” Nov. 2021, Accessed: Jun. 21, 2025. [Online]. Available: http://arxiv.org/abs/2103.14749

[30] “Home - UCI Machine Learning Repository.” Accessed: Jun. 24, 2025. [Online]. Available: https://archive.ics.uci.edu/

[31] “OpenML.” Accessed: Jun. 21, 2025. [Online]. Available: https://www.openml.org/search?type=data&sort=runs&id=182&status=active

[32] “sklearn.datasets — scikit-learn 1.7.0 documentation.” Accessed: Jun. 24, 2025. [Online]. Available: https://scikit-learn.org/stable/api/sklearn.datasets.html

[33] “Regulation - 2016/679 - EN - gdpr - EUR-Lex.” Accessed: Jun. 25, 2025. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[34] R. D. Peng, “Reproducible Research in Computational Science,” *Science (1979)*, vol. 334, no. 6060, pp. 1226–1227, Dec. 2011, doi: 10.1126/SCIENCE.1213847.

[35] “Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. MIT Press. - References - Scientific Research Publishing.” Accessed: Jun. 25, 2025. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=3614746

[36] “Ethics guidelines for trustworthy AI | Shaping Europe’s digital future.” Accessed: Jun. 25, 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[37] “Balance Scale - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/12/balance+scale

[38] “Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[39] “Car Evaluation - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/19/car+evaluation

[40] “Dermatology - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/33/dermatology

[41] “Glass Identification - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/42/glass+identification

[42] “Iris - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/53/iris

[43] “Phishing Websites - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/327/phishing+websites

[44] “Seeds - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/236/seeds

[45] “User Knowledge Modeling - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/257/user+knowledge+modeling

[46] “Vertebral Column - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/212/vertebral+column

[47] “Wine - UCI Machine Learning Repository.” Accessed: Jun. 21, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/109/wine