COMPARING CLASSIFIERS IN THE PRESENCE OF ERRORS IN TRUE LABEL ASSIGNMENT IN MEDICAL DATASETS

Vishwa Vallabh Angampally and Eugene Pinsky

Department of Computer Science, Metropolitan College Boston University, 1010 Commonwealth Avenue Boston, MA 02215, USA

ABSTRACT

We often rely on human experts to assign true labels in medical datasets, which may not be 100% accurate. We investigate the impact of labeling errors on machine-learning classi- fiers applied to medical datasets. By introducing symmetric errors from 0% to 40% in True labels— simulating errors of true labels assignment by experts, inter-observer variability, and automated annotation - we evaluate the impact of such errors in binary classification for several well-known medical datasets using traditional machine learning models and metrics. Although all models experience degradation as errors increase, simpler, well-regularized methods such as Logistic Regression and SVM decline more gracefully. Our results underscore the necessity for improved data curation and error-aware training strategies in medical AI, ultimately guiding the selection of robust algorithms that maintain reliability under imperfect real-world conditions.

KEYWORDS

Labeling Errors, Medical Datasets, Classifier Robustness, Data Imbalance, Machine-Learning, Performance Evaluation, Diagnostic Prediction.

1. INTRODUCTION

The reliability of medical datasets is fundamentally challenged by labeling errors arising from multiple sources, including human misinterpretation, inter-observer variability, and limitations of automated annotation systems. In clinical practice, even expert practitioners can disagree on diagnoses due to ambiguous symptom presentations or limitations in imaging and laboratory test interpretations ([1], [2]), potentially undermining the performance of supervised learning models. Such inconsistencies are especially problematic in high-stakes decision-making, where minor labeling errors may result in significant diagnostic or prognostic inaccuracies.

Prior research has shown that even modest label noise can significantly degrade the performance of supervised learning models ([3], [4]) underscoring the need for noise-aware data curation and algorithmic robustness in medical AI. In our study, we focus on traditional machine learning classifiers to assess the impact of label noise on predictive performance across multiple clinical prediction tasks (sepsis prediction, breast cancer prognosis, stroke detection, heart disease diagnosis, and diabetes prediction). To replicate common labeling errors, we simulate random symmetric noise ranging from 0% to 40%, increasing in 5% increments. Al- though symmetric noise injection, in which labels are flipped uniformly at random, is a common method for assessing model robustness ([5], [6]), it does not always capture the complex, often systematic nature of mislabeling in real clinical settings. In reality, label

DOI:10.5121/mlaij.2025.12201

noise in medicine is class-dependent; for example, false negatives may be more common than false positives due to a clinician's tendency to overlook mild signs of a disease ([7] [8]).We analyze six traditional machine learning classifiers, including Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), Naïve Bayes (NB), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). These algorithms were chosen because they are well- established, interpretable, and generally perform well on structured medical data, especially when using limited sample sizes. In many healthcare classification tasks using structured data, simple models like Logistic Regression (LR) often achieve results that are comparable to com- plex models [9]. For several reasons, we intentionally opted not to use deep neural networks (DNNs) in our evaluation. First, many of the datasets under consideration have limited sample sizes, making them unsuitable for training high-capacity DNN models, which are known to be especially prone to overfitting when subjected to noisy labels because of their high capacity to memorize incorrect labels. ([10], [11], [12]). Second, the interpretability of traditional models such as Logistic Regression and Support Vector Machines allows for a clearer understanding of how label noise impact model decision boundaries a crucial factor when translating findings into clinical practice [13]. Although advanced loss-correction and noise-adaptive training strategies have shown promise [14], they often require larger, well-curated datasets and extensive computational resources, beyond the scope of this investigation which will be a natural extension of this study in future works. Given our goal of analyzing the impact of label errors without additional confounding factors, we opted for classical algorithms, which are faster to train and easier to interpret. This approach aligns with prior efforts that focuses on label noise in smaller datasets, which frequently use traditional classifiers for baseline comparisons [9].

Our experiments systematically quantify degradation in standard metrics (accuracy, TPR, TNR, F1-score, and PPV) as noise increases. Though random symmetric noise does not entirely replicate the complex nature of real clinical mislabeling, In future work, we plan to explore class- dependent and systematic noise models, including deep learning architectures, aiming further to enhance the reliability and robustness of medical predictive models.

1.1. Review of Prior Literature

Label noise—errors in target annotations—has been recognized as a key challenge in machine learning, with early studies demonstrating that even modest mislabeling can lead to substan- tial performance degradation. For instance, Brodley et al. [3] showed that mislabeled training data undermines classifier reliability, exposing models' vulnerability to even modest labeling errors. In a related study, Zhu et al. [4] quantitatively differentiated between class noise (in- correct labels) and attribute noise (errors in feature measurements), a distinction that has informed many subsequent noise-modeling approaches. Later research introduced the concept of class-conditional noise, emphasizing that errors affecting minority classes can be particularly damaging. In one stroke dataset study, flipping the rare positive cases led to disproportionately large drops in sensitivity [6].

As the field evolved, comprehensive surveys like the one by Freney et al. [15] provided de-tailed taxonomies of noise types and reviewed a variety of methods for detecting and mitigating label noise. Building on these insights, robust techniques emerged to address noisy labels in a targeted manner. For instance, Bootkrajang et al. [16] advanced noise-robust logistic regression models that help prevent overfitting to erroneous labels—a key advancement when expert an- notations are both costly and prone to error. More recently, Northcutt et al. [17] introduced the Confident Learning framework, which quantitatively estimates label uncertainty and provides a principled approach for identifying suspect labels in large datasets. Such strategies are espe- cially critical in medical contexts, where

misdiagnoses or ambiguous labels can not only reduce model performance but also lead to dire clinical consequences.

With the advent of deep learning, additional challenges have emerged. Deep neural networks, while powerful, are particularly susceptible to overfitting on noisy data. This has led to the development of specialized methods such as noise adaptation layers [10] and co-teaching frameworks [18], which leverage multiple models simultaneously to filter out noisy instances. Furthermore, recent techniques employing self-attention and self-supervised learning have been shown to improve the robustness of models in the presence of label noise [2]. A recent com- prehensive survey by Shi et al. [19] highlights a diverse array of label-noise handling strategies in deep learning for medical image analysis, reinforcing the necessity for robust, noise-aware training techniques even within classical machine learning settings.

Notably, while extensive research has addressed label noise in computer vision and general machine learning, the unique challenges inherent to medical datasets remain largely underexplored. Unlike prior studies that often focus narrowly on a single dataset or disease, our research systematically explores the impact of label noise across multiple diverse medical datasets. Furthermore, in contrast to studies predominantly centered on deep neural networks,we comprehensively evaluate traditional, interpretable classifiers (Logistic Regression, Decision Trees, Random Forests, Naïve Bayes, KNN, and SVM) methods particularly valuable in clinical settings due to their interpretability and computational efficiency. Additionally, we critically examine the interaction between class imbalance and label noise. By explicitly bridging theory and practical implementation, our paper recommends choosing robust classifiers and noise-aware training strategies, ultimately enhancing the reliability and diagnostic precision of real-world medical machine learning applications.

In practical system designs, while it is often useful to isolate and study the effects of noise under controlled conditions, real-world applications can benefit from integrating a preliminary noise-detection step. For example, removing outlier instances that appear mislabeled before applying oversampling or other data augmentation strategies can further enhance model per- formance ([2], [20]). This holistic approach to managing label noise is essential to ensure that machine learning models in sensitive domains such as healthcare remain both accurate and reliable.

1.2. Datasets

To test robustness in the presence of true labeling errors, we selected five public datasets that span different clinical tasks, disease prevalence, and feature types. This diversity ensures that our findings reflect various real-world challenges, from large-scale electronic health record (EHR) data to smaller genomic studies. Moreover, each dataset represents a clinically meaningful prediction task—diagnosis or prognosis—that aligns with practical needs in healthcare. Below, we describe each dataset in detail

- **Sepsis Prediction (MIMIC-IV)**: This dataset is derived from the publicly available MIMIC-IV database, which comprises de-identified electronic health records from critical care units. We extracted approximately 94,458 samples of adult ICU stays from this extensive resource. Each record includes diverse clinical variables such as demographic data (age, gender), vital signs (e.g., blood pressure, respiratory rate, temperature), laboratory measurements (e.g., lactate levels, white blood cell count), and documented comorbidities. The goal is to predict the onset of sepsis—a severe and life-threatening response to infec-

tion. The labeled outcome indicates whether sepsis eventually develops, making it a binary classification task. This dataset is highly imbalanced, with sepsis occurring rela- tively infrequently. The possibility of label noise arises from diagnostic uncertainties and inconsistencies in the exact timing or severity level at which a clinician identifies sepsis, all of which can lead to mislabeling in the electronic health record ([21], [22]).

- **GSE2034 (Breast Cancer Prognosis)**: GSE2034 is a microarray-based gene expres- sion dataset focusing on node-negative breast cancer patients, comprising 286 samples in total. The dataset measures expression levels of thousands of genes, supplemented by basic clinical and demographic factors. The binary target class indicates whether a patient experiences distant metastasis (i.e., cancer recurrence) within a certain follow-up window. The data present an inherent class imbalance because the positive (recurrence) class is smaller. Furthermore, the high-dimensional nature of gene expression profiles means each sample has numerous features capturing complex biological signals. This complexity, coupled with limited sample size, makes any degree of label noise particularly consequential [23].

- **Stroke Prediction**: This dataset consists of 5,110 patient records compiled from various clinical and demographic sources, each containing 12 features. Common predictors include hypertension status, smoking history, average glucose levels, BMI, and other health indica- tors associated with cerebrovascular events. The binary outcome denotes whether or not a patient suffered a stroke, capturing both ischemic and hemorrhagic types. Although mod- erately sized, this dataset has a distinct positive-class rarity—fewer than 5% of patients experienced a stroke—which makes the prediction problem significantly imbalanced [24].

- **Heart Disease Diagnosis (UCI dataset)**: This is a classic dataset from the UCI Machine Learning Repository, comprising 303 patient records related to coronary artery disease. Each record is accompanied by 13 clinical attributes, such as chest pain type, resting ECG results, serum cholesterol, and maximum heart rate achieved during physical exercise. The objective is to distinguish between individuals who do or do not have heart disease, making it a relatively balanced classification problem (roughly equal numbers of positive and negative cases). Despite its smaller size, this dataset remains widely used in benchmarks due to the relevance of features, their interpretability, and the simplicity of data preprocessing [25].

- **Diabetes (Pima Indians) Prediction (UCI dataset)**: Another dataset from the UCI repository, this collection includes 768 samples of Pima Indian women, each with 8 numeric features such as plasma glucose concentration, diastolic blood pressure, body mass index, diabetes pedigree function, and age. The objective is to predict the onset of type 2 diabetes, leading to a binary classification of diabetic versus non-diabetic status. Although the class distribution is not extremely skewed, roughly a third of the samples are positive, which is enough to introduce moderate imbalance [26].

1.3. Classifiers

We evaluated the following models:

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Naïve-Bayes (NB)
- K-Nearest Neighbors (KNN)

- Support Vector Machine (SVM)

2. EXPERIMENTAL SETUP

We imputed missing numeric features via median or mean, one-hot encoded categorical values, and standardized continuous variables. A stratified 50% train-test split was repeated multiple times to reduce variance in performance.

2.1. Label Noise Simulation

To study the impact of incorrect labels, Following methodologies similar to [5] and [6],we injected synthetic noise into the training labels, by randomly flipping the ground truth labels of a fixed percentage of samples.

Specifically, we applied symmetric label noise at varying rates to only the training set. Symmetric noise (also known as random noise) means any given training label has a fixed probability p of being switched with an incorrect label chosen uniformly at random from the other classes [27]. This approach, while not capturing class-dependent noise patterns common in medical settings, provides a controlled baseline for comparing model robustness [5]. Future work will extend this to class-biased noise models (e.g., higher mislabeling rates for minority classes). While asymmetric noise (class biased) can occur in real clinical settings (owing to systematic biases or certain classes being more prone to error) [27], we focused on symmetric noise for this initial analysis to isolate the impact of random errors. This assumption of noise being random and unbiased is commonly used in label noise research [28].

Noise levels ranged from 0% (i.e., clean labels) up to 40%, increasing in 5% increments. Each noise level experiment was repeated multiple times to account for randomness. In our approach, we used different random seeds for the noise injection process and repeated each training/testing experiment N=10 times. This method calculates an average performance and a measure of variability for each classifier at each noise level. By varying the random seed, we effectively simulate different "realizations" of label noise, improving the robustness of our findings. To ensure reproducibility, all processes (such as label flipping and data shuffling) were controlled on recorded seed values.

After noise injection, the training sets were balanced using different sampling strategies based on dataset characteristics. Specifically, random oversampling was applied for the genomics (GSE2034) and stroke datasets. In contrast undersampling was employed for the sepsis dataset to manage its large sample size and avoid overwhelming the minority class [20].

2.2. Evaluation Metrics

We track five key metrics:

- **Accuracy**: Proportion of correct predictions.
- **True Positive Rate (TPR)**: Fraction of positive cases correctly identified.
- **True Negative Rate (TNR)**: Fraction of negative cases correctly identified.
- **Positive Predictive Value (PPV)**: Added to indicate the precision or the proportion of predicted positives that are true positives.
- **F1 Score**: Harmonic mean of precision (PPV) and sensitivity (TPR), providing a bal-

anced measure of model performance.

- In addition to reporting standard performance metrics, we conducted a statistical analysis calculating the slope of metric degradation between noise levels (0% to 40%) for each classifier. For each dataset, a linear regression was fitted to the metric values, and the resulting slope indicates how quickly the accuracy of a model declines as the noise of the labels increases. A smaller absolute slope value signifies greater robustness.

3. KEY OBSERVATIONS

3.1. MIMIC-IV (Sepsis) Dataset

Class Distribution:

- Class 0: 84,826
- Class 1: 9,632

Table 1: Performance of six classification models for sepsis detection on MIMIC-IV under increasing levels of label noise (0%, 20%, and 40%).

	Accuracy			TPR			TNR			PPV			F1 Score		
Model	**0%**	**20%**	**40%**	**0%**	**20%**	**40%**	**0%**	**20%**	**40%**	**0%**	**20%**	**40%**	**0%**	**20%**	**40%**
LR	71.1	62.9	59.0	73.3	81.7	78.4	70.8	60.4	56.3	25.5	22.0	19.7	37.8	34.6	31.5
DT	66.7	53.4	51.0	65.9	61.1	53.9	66.8	52.4	50.6	21.3	14.9	13.0	32.2	24.0	20.9
RF	74.8	66.6	57.7	78.4	78.4	65.3	74.3	65.0	56.6	29.5	23.6	17.1	42.8	36.2	27.1
NB	69.4	50.9	40.6	48.0	58.4	66.9	72.3	49.9	37.0	19.1	13.8	12.7	27.3	22.3	21.3
KNN	70.6	56.7	51.8	66.4	61.3	54.0	71.2	56.1	51.5	24.0	16.0	13.3	35.2	25.4	21.3
SVM	71.4	70.0	63.3	78.1	76.0	68.9	70.4	69.2	62.5	26.5	25.4	20.1	39.5	38.1	31.1

Key Observations:

- SVM starts at 71.4% accuracy and only drops 8.1 points (to 63.3%) under 40% noise, the smallest decline among all models. Both TPR (78.1% → 68.9%) and TNR (70.4% → 62.5%) decrease gently, indicating balanced resilience.
- KNN and RF are notably more sensitive to label noise, each losing 17+ points of accuracy by 40% noise.
- LR's TPR increases from 73.3% to 78.4%, while TNR steadily declines (70.8% → 56.3%). In contrast, SVM's TPR/TNR both show gentler decreases.
- Naive Bayes shows large swings in TPR/TNR and ends up with the largest overall drop in accuracy (down to 40.6% from 69.4%).
- Although RF had the highest clean-data accuracy (74.8%), its advantage diminished under noise; SVM, despite a lower baseline (71.40% accuracy), maintained 63.30% accuracy with relatively balanced TPR and TNR at high noise.
- In the sepsis prediction task, SVM again outperforms with the lowest slope (–0.001843), followed by Logistic Regression at –0.002700, highlighting its resilience in noisy conditions. These patterns echo prior findings that high-capacity models can overfit to noisy labels more readily [29].

3.2. Genome Breast Cancer (GSE2034) Dataset

Class Distribution:

- Class 0: 217
- Class 1: 69

Table 2: Performance of six classification models for breast cancer detection on GSE2034 under increasing levels of label noise (0%, 20%, and 40%).

Accuracy				TPR			TNR			PPV			F1 Score		
Model	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%
LR	76.7	70.3	56.2	23.0	32.7	44.9	93.6	82.2	59.8	53.4	36.6	26.1	32.1	34.5	32.9
DT	65.0	57.6	51.7	28.3	41.7	45.5	76.5	62.6	53.6	27.5	26.0	23.6	27.9	32.0	31.0
RF	76.2	71.8	57.5	10.5	15.5	35.9	96.9	89.6	64.3	51.4	31.9	24.1	17.4	20.8	28.8
NB	72.7	67.8	55.9	33.2	33.8	44.6	85.1	78.6	59.5	41.3	33.2	25.8	36.8	33.5	32.7
KNN	65.1	58.8	54.2	30.3	39.7	42.3	76.1	64.9	58.0	28.6	26.3	24.1	29.4	31.6	30.6
SVM	64.3	62.9	54.3	63.0	54.5	45.5	64.7	65.5	57.1	36.0	33.2	25.1	45.7	41.2	32.3

Key Observations:

~

Key Observations:

- SVM is the most robust in accuracy, degrading by only ˜10 points.
- KNN also degrades relatively little (˜11 points), surprisingly more stable here than in some other datasets.
- LR and RF, which start with the highest accuracies (˜76% but low TPR) experience fairly large drops (˜20 and 19 points, respectively). Their TNR plummets more severely than most, showing that they become more susceptible to false positives under label noise.
- SVM achieves the best performance with a slope of -0.002581, while logistic regression suffers the steepest degradation (–0.004916) followed by Random Forest (-0.004720), in- dicating that SVM retains precision more effectively with increasing Label Noise in this highly dimensional dataset.

3.3. Stroke Dataset

Class Distribution:

- Class 0: 4,860
- Class 1: 249

Table 3: Performance of six classification models for stroke prediction under increasing levels of label noise (0%, 20%, and 40%).

	Accuracy			TPR			TNR			PPV			F1 Score		
Model	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%
LR	74.09	59.87	52.03	79.52	84.88	56.80	73.81	58.59	51.78	0.135	0.095	0.057	0.231	0.171	0.104
DT	91.69	74.62	57.55	12.48	29.52	43.36	95.77	76.94	58.28	0.132	0.062	0.051	0.128	0.102	0.091
RF	94.26	85.79	64.63	2.64	17.20	38.08	98.97	89.32	66.00	0.117	0.077	0.054	0.043	0.106	0.095
NB	26.87	59.74	44.83	99.28	76.16	69.12	23.15	58.90	43.58	0.062	0.087	0.059	0.117	0.155	0.109
KNN	85.77	56.51	52.13	25.76	50.64	50.56	88.86	56.81	52.21	0.106	0.057	0.052	0.150	0.102	0.094
SVM	78.35	62.04	53.71	56.72	63.28	52.88	79.46	61.98	53.75	0.124	0.079	0.055	0.204	0.140	0.100

Key Observations:

- Most models degrade sharply, losing 22–34 points of accuracy by 40% noise. This is one of the largest noise impacts across datasets.
- Random Forest holds the top accuracy across all noise levels but suffers a 29.6-point drop (94.26% → 64.63%), revealing notable sensitivity to high label noise.
- LR has the smallest negative drop among the high-accuracy models (~22 points, whereas SVM, KNN, and RF all lose ~30 points from their 0% baselines.
- DT sees the largest absolute decline (34 points). It starts at 91.69% and drops to 57.55%.
- Although overall degradation is higher in this dataset, Logistic Regression (–0.005020) and SVM (–0.005907) maintain relatively lower slopes compared to Random Forest (–0.007474) and KNN (–0.007522).

3.4. Diabetes Dataset

Class Distribution:

- Class 0: 500
- Class 1: 268

Table 4: Performance of six classification models for diabetes prediction under increasing levels of label noise (0%, 20%, and 40%).

	Accuracy			TPR			TNR			PPV			F1 Score		
Model	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%
LR	0.771	0.762	0.700	0.587	0.570	0.523	0.868	0.864	0.795	0.703	0.692	0.577	0.639	0.625	0.549
DT	0.698	0.596	0.537	0.593	0.531	0.522	0.753	0.631	0.544	0.563	0.436	0.380	0.576	0.479	0.440
RF	0.757	0.723	0.600	0.603	0.569	0.497	0.840	0.806	0.654	0.669	0.611	0.435	0.634	0.588	0.464
NB	0.748	0.743	0.691	0.600	0.581	0.522	0.829	0.830	0.781	0.653	0.647	0.562	0.625	0.612	0.540
KNN	0.727	0.670	0.567	0.582	0.548	0.502	0.805	0.735	0.601	0.615	0.525	0.403	0.599	0.536	0.447
SVM	0.760	0.750	0.673	0.560	0.531	0.463	0.867	0.867	0.785	0.693	0.682	0.535	0.620	0.597	0.496

Key Observations:

- Naive Bayes emerges as the most stable when looking at the magnitude of performance changes, especially in accuracy (74.8% → 69.1%, a ~5.7-point drop) and PPV (65.3% → 56.2%). Its TPR/TNR balance shifts less drastically than other models.
- LR and SVM indeed degrade more gradually (7-9 points) compared to KNN and

RF (15+).

- DT and KNN show marked vulnerability to label noise, falling below 60% accuracy at 20% noise and ending near the mid-50% range at 40% noise. TNR, in particular, collapses quickly for DT.
- RF starts off strong but ends up at 60.0% accuracy by 40% noise. While it still outperforms DT and KNN in raw accuracy, its TNR and PPV have large absolute drops, indicating a bigger shift away from its conservative baseline.
- Logistic Regression (–0.001427) and SVM (–0.001740) show notably less steep declines in accuracy than Random Forest (–0.003663), Decision Tree (–0.003980), and KNN (–0.004097), suggesting higher stability under noise.

3.5. Heart Disease Dataset

Class Distribution:

- Class 0: 160
- Class 1: 137

Table 5: Performance of six classification models for heart disease prediction under increasing levels of label noise (0%, 20%, and 40%).

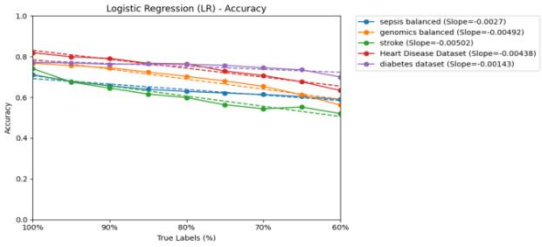
	Accuracy			TPR			TNR			PPV			F1 Score		
Model	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%
SVM	0.813	0.781	0.662	0.764	0.730	0.659	0.856	0.824	0.665	0.821	0.781	0.629	0.791	0.755	0.644
LR	0.819	0.764	0.634	0.777	0.722	0.652	0.856	0.800	0.618	0.823	0.757	0.595	0.799	0.739	0.622
KNN	0.795	0.746	0.612	0.709	0.674	0.606	0.869	0.809	0.618	0.823	0.752	0.577	0.762	0.711	0.591
DT	0.713	0.630	0.538	0.681	0.593	0.522	0.740	0.661	0.552	0.693	0.601	0.501	0.687	0.597	0.511
NB	0.699	0.593	0.561	0.593	0.632	0.657	0.791	0.560	0.479	0.710	0.553	0.521	0.646	0.590	0.581
RF	0.789	0.760	0.619	0.719	0.720	0.616	0.850	0.794	0.623	0.805	0.751	0.585	0.760	0.735	0.600

Key Observations:

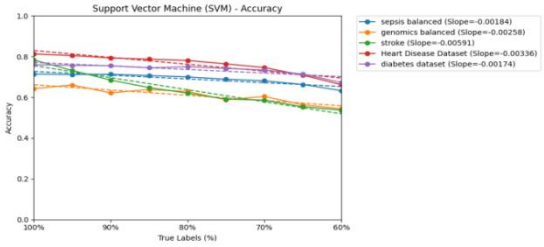
- SVM distinguishes itself with high overall performance and a smaller drop than many models, leading at both the mid (20%) and high (40%) noise points in accuracy and F1.
- LR, KNN, and RF all experience moderate-to-large overall declines, though each retains a reasonable balance between TPR and TNR until very high noise.
- DT suffers more considerably, falling below 54% accuracy at 40% noise with the lowest final F1
- Within the Heart Disease dataset, the Support Vector Machine (SVM) exhibits the slowest rate of accuracy decline (following NB) with a slope of –0.003363. In comparison, Logistic Regression, Decision Trees, Random Forests, and K-Nearest Neighbors have steeper slopes (–0.004383, –0.004290, –0.004453, and –0.004630 respectively), indicating that SVM is less adversely affected by increasing label noise.

Model Wise Accuracy and Slope Comparison

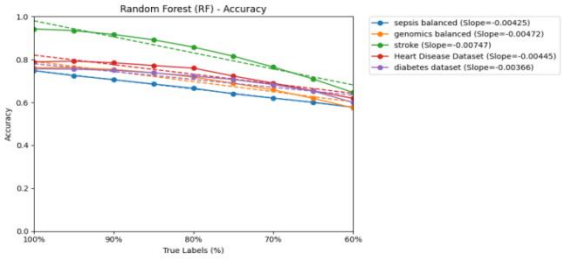
Logistic Regression - Accuracy



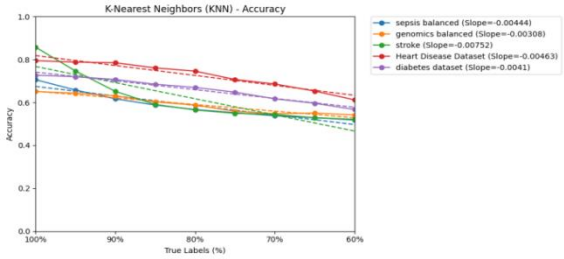
Support Vector Machine - Accuracy



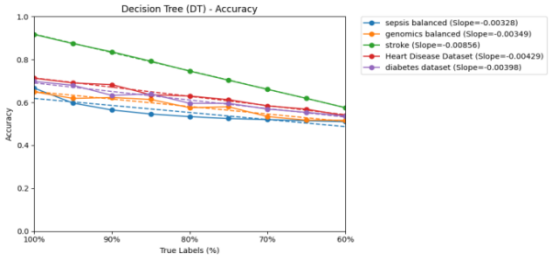
Random Forest - Accuracy



K-Nearest Neighbours - Accuracy



Decision Tree - Accuracy



Naive Bayes - Accuracy

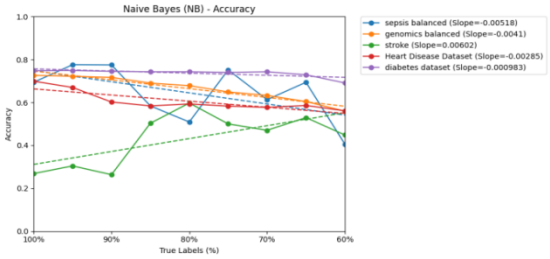


Figure 1: Model wise Accuracy and Slope Comparison across the 5 Datasets

Table 6: Model Degradation slopes

Sepsis (MIMIC-IV)	Genomics (GSE2034)	Stroke	Diabetes	Heart Disease
SVM: -0.001843	SVM: -0.002581	NB: +0.006016	NB: -0.000983	NB: -0.00285
LR: -0.002700	KNN: -0.003084	LR: -0.005020	LR: -0.001427	SVM: -0.003363
DT: -0.003283	DT: -0.003492	SVM: -0.005907	SVM: -0.001740	DT: -0.004290
RF: -0.004247	NB: -0.004101	RF: -0.007474	RF: -0.003663	LR: -0.004383
KNN: -0.004443	RF: -0.004720	KNN: -0.007522	DT: -0.003980	RF: -0.004453
NB: -0.005180	LR: -0.004916	DT: -0.008564	KNN: -0.004097	KNN: -0.004630

4. DISCUSSION

4.1. Impact of Label Noise

Our findings underscore several patterns regarding label noise tolerance in medical dataset contexts. First, we observed clear tipping points beyond which model performance degrades rapidly. In most datasets, performance was relatively stable up to about 10% noise (sometimes 15%), with only small increases in error. Beyond this threshold, additional noise led to more pronounced declines. For example, in the heart disease dataset a noticeable inflection occurred at 15% noise; in the gene expression data, error rates accelerated after 10% label noise. These tipping points likely correspond to the point at which the noise starts overwhelming the true signal in the data. Up to a low noise level, models can often identify and largely ignore a few mislabeled outliers (treating them as noise in the error term). But once a critical mass of labels is wrong, the algorithms begin to learn incorrect decision boundaries. This critical noise level will depend on the dataset's complexity and redundancy of predictive features [15], [6], [7]). For GSE2034 (high complexity, limited samples), the tipping point was quite low; for a more redun- dant dataset like heart disease (many patients, correlated features like chest pain and ECG), the models withstood higher noise before breaking down. In practical terms, this suggests that if a medical dataset is suspected to have label error rates beyond 10-15%, one should be prepared for significant performance issues or should apply noise-robust training techniques. Sensitiv- ity, critical for detecting rare positive cases, is particularly affected—especially in imbalanced datasets such as stroke prediction. These findings echo prior research ([15], [6]) which found that even modest mislabeling can substantially confuse classifiers when the positive class is scarce. The Observed rapid decline in Performance with increasing label noise is earlier findings that highlight a tipping point beyond which the true signals are overwhelmed ([6], [7]). Furthermore, Khanal et al. [8] provide evidence that when label noise is class-dependent—especially in sce- narios with high inter-class visual similarity—it can adversely affect not only the noisy classes but also propagate errors to otherwise clean classes

4.2. Model-Specific Robustness

In noise-free or low-noise conditions, more complex models (e.g., random forests) often yield higher accuracy – they can capture nonlinear patterns and interactions that simpler models miss. However, as the noise grew, these complex models typically suffered larger performance hits, allowing simpler models to overtake them. After considering all the metrics across the 5 datasets we see:

1. Logistic Regression (LR): LR exhibited one of the most gradual decline across noise levels. Its regularized linear structure prevented overfitting, thereby preserving a relatively bal- anced sensitivity and specificity—even as overall accuracy dropped. Its linear form and convex training objective likely contributed to stable performance – it cannot memorize outliers the way a non-parametric method can. This gentle decline under noisy conditions may make LR a suitable choice when we are unsure of the data quality. Its robustness can be demonstrated by observing its slope values across multiple datasets. For exam- ple, in the Diabetes dataset, LR's accuracy slope is –0.001427, meaning its performance barely decreases even with rising noise. In contrast, in the Genomics (GSE2034) dataset, LR's slope is –0.004916, suggesting that high-dimensional, small-sample conditions make it more sensitive. In the heart disease dataset, LR's slope of –0.004383 places it in the mid-range; however, in the Sepsis

dataset, LR's slope of –0.002700 highlights its strong robustness even in imbalanced settings. In the challenging Stroke dataset, LR's slope is –0.005020—still more gradual than the declines seen for methods such as Random Forest, KNN, and DT. Overall, LR's consistently low slopes (especially in Diabetes and Sepsis) provide strong evidence of its robustness under label noise. This superior robustness of simpler models, such as logistic regression in noisy environments, is in line with earlier works on noise robust methods ([16], [29]).

2. Support Vector Machine (SVM): Although SVM achieved high baseline accuracy, its re- liance on support vectors made it more vulnerable under noise. In the heart disease dataset, its slope of –0.003363 is lower than that of Random Forest (–0.004453) and KNN (–0.004630), indicating less sensitivity to noise. In the Diabetes dataset, SVM's
   –0.001740 slope is competitive, while in the Genomics dataset it achieves a top slope of
   –0.002581. Furthermore, in the Sepsis dataset, SVM records the best performance with a slope of –0.001843. Even in the Stroke dataset—where slopes are overall higher—SVM's
   –0.005907 remains lower than those of RF, KNN, and DT. These results indicate that SVM is generally more effective at maintaining accuracy as label noise increases. With strong regularization, SVM could potentially improve, but our default RBF SVM declined faster than LR in some settings [13].

3. Random Forest (RF): RF started with high accuracies, but its advantage eroded as noise increased. Its ensemble averaging helps, but when many trees are affected by noise, both sensitivity and specificity suffer. For example, in the Diabetes dataset its slope is –0.003663 and in the Heart Disease dataset it is –0.004453. In the Genomics dataset, RF's slope is –0.004720, and in the Sepsis dataset, it reaches
   –0.004247. These steeper slopes (Compared to LR and SVM) suggest that the ensemble benefits of RF are compro- mised when many trees are impacted by mislabeled data, resulting in reduced sensitivity and specificity. The averaging effect in RF reduces variance and dilutes the impact of any single noisy instance. Yet, RF can still overfit if many trees see the same mislabeled points, especially when the noise is widespread (Falloff point at 20-25% across datasets).

4. K-Nearest Neighbors (KNN) and Decision Tree (DT): Both were highly sensitive to noise. In the Diabetes dataset, KNN's slope is –0.004097 and DT's is –0.003980. In the Heart Disease dataset, their slopes are –0.004630 (KNN) and –0.004290 (DT). In the Genomics dataset, KNN and DT have slopes of –0.003084 and –0.003492, respectively. Furthermore, in the Sepsis dataset KNN's slope is –0.004443 and DT's is –0.003283, and in the Stroke dataset, the declines are even steeper at –0.007522 for KNN and –0.008564 for DT. These values indicate that both methods rapidly lose accuracy as label noise increases, making them less reliable when data quality is questionable. KNN, lacking any built-in noise filtering, and DT, with its greedy splits, suffered the most dramatic performance drops. In these cases, sensitivity and specificity declined to near-chance levels under high noise. For Decision Trees Pruning or limiting depth helps but doesn't solve the problem. K-Nearest Neighbors also struggle with noise, as incorrect neighbors can influence the majority vote. A large k reduces sensitivity but dilutes the local signal. While robust KNN variants exist, we used the standard version. Our results indicate KNN is unreliable with even modest label errors, a key point for medical settings where it is sometimes used for its simplicity and interpretability [4].

5. Naïve Bayes (NB) exhibited dataset-dependent behavior. In stroke prediction, it maintained higher sensitivity at moderate noise due to tuned probabilistic thresholds for posi- tives, while other models became biased toward negatives [30]. However, NB's specificity suffered, and in general, its accuracy was middling. NB's strong assumptions (feature independence) act as a double-edged sword: they prevent it from fitting complex patterns (which is good under noise), but they also limit its peak accuracy on clean data.

These cross-over points highlight a practical consideration: when data quality is high, one might prefer a powerful model to maximize accuracy, but if even moderate label noise is present, a simpler approach could actually generalize better on clean test data. This reinforces the importance of diagnosing dataset quality; a clinician or data scientist should not automatically choose the model with the best in-sample performance without considering the potential errors in labels.

4.3. Role of Class Imbalance

The chosen sampling strategies effectively balanced the training data. For the genomics (GSE2034) and stroke datasets, random oversampling was used, which can replicate mislabeled instances and amplify noise effects [20]. In contrast, for the sepsis dataset, under sampling was adopted due to its large size, and extreme imbalance. This method reduced training time and prevented the majority class from dominating the learning process. However, undersampling may impact the noise distribution differently from oversampling by potentially discarding some informa- tive examples, which could affect the model's ability to generalize [31]. Although oversampling improved baseline sensitivity, it did not prevent substantial declines in sensitivity as noise in- creased. Future work will explore alternative strategies, such as synthetic data generation (e.g., SMOTE) or cost-sensitive learning, to better balance these challenges.

4.4. Practical Implications

Below are the key practical implications drawn from our work:

- Our study's real-world implications include the importance of data quality in medical AI. Medical labels (diagnoses, outcomes, etc.) can often be noisy in practice due to recording errors, inter-observer variability, or ambiguous conditions [15]. Our results quantifiably show that even modest rates of label noise (10-15%) can significantly effect a model's performance metrics like sensitivity or specificity which is especially concerning for high-stakes applications like sepsis early-warning systems. This finding emphasizes the need for rigorous data curation via expert reviews, consensus-driven labeling, and the implementation of standardized annotation protocols with periodic audits [6].

- When using secondary EHR data, researchers should be aware of potential mislabeling and consider techniques to handle it, such as robust training algorithms or human-in-the-loop correction for outliers. Techniques such as cost-sensitive learning [29], noise-robust loss functions [16], or even recent frameworks like Confident Learning [17] have been shown to improve performance on contaminated datasets. Moreover, recent advances in active label cleaning—such as the approach proposed in [32] offer a promising avenue for iteratively refining datasets by prioritizing re-annotation of the most uncertain samples under limited resource conditions

- In scenarios where training data are inherently noisy, simpler models that avoid overfitting to outliers—such as Logistic Regression and Support Vector Machines—may be preferable, as they have demonstrated a more gradual performance decline compared to more complex alternatives.

- In clinical environments, the robustness of a model under imperfect data conditions is crucial. Regulatory bodies and practitioners may favor models that demonstrate stability and reliable performance over those that only excel on meticulously curated datasets. Pro- viding quantitative benchmarks on noise tolerance can support evidence-based decision- making and encourage deployment strategies that prioritize resilience to data imperfec- tions, even at the cost of a modest reduction in peak performance [13].

4.5. Limitations

There are a few Limitations to our study

- First, the noise we injected is *random (symmetric) noise*. In reality, label noise in medicine is often *systematic* or *class dependent*. For example, perhaps false negatives are more common than false positives (a doctor might miss a condition rather than label a healthy person as ill) [15], [6]. Our experiments did not specifically simulate class-biased noise, which can have different effects (often more detrimental, since it effectively changes class prevalence and confuses the learner) . Future work could introduce noise that preferen- tially flips positives to negatives or vice versa to mimic real error tendencies.

- Second, our use of oversampling means the effective training size increases for minority classes. While this helped us analyze sensitivity, it also means the training distributions at different noise levels weren't strictly identical in terms of sample count [20]. We attempted to control this by keeping the oversampling procedure consistent, but there is a possibility that some observed effects are partly due to interactions of noise with oversampling. Using alternative imbalance methods (or evaluating on the imbalanced test set) might shed more light here. Combining noise filtering with oversampling (e.g., SMOTE+ENN) could further mitigate the adverse effects of replicated noise [33].
- Third, we focused on traditional machine learning models. We did not include neural network models, which are known to be powerful but also especially prone to label noise memorization [10]. The choice was partly due to the limited sample size of some datasets (deep learning would risk overfitting anyway on such small data) and to maintain inter- pretability of results. Although this choice was motivated by concerns regarding dataset size and interpretability, it restricts the generalizability of our findings. However, deep learning is increasingly applied to larger medical datasets (like MIMIC-IV); exploring noise impact on neural networks with techniques such as dropout, noise-robust loss functions, or semi-supervised learning would be a valuable extension of this study we plan on pursuing in the future works.

- We used default or limited hyperparameter settings. Since models like RBF SVMs and Random Forests are sensitive to tuning, our results might partly reflect suboptimal con- figurations rather than true robustness. Future work should include extensive sensitivity analysis and hyperparameter optimization across noise levels.

- Lastly, our analysis treated each dataset in isolation. One could also consider model transferability under noise: e.g., train on a noisy dataset and test on a different (clean) dataset for the same task – how well does the model generalize? This could reflect scenarios like training on one hospital's noisy labels and testing on another's validated registry. We did not explore that cross-dataset transfer.

4.6. Noise Mitigation Strategies

Our experiment was intentionally designed to investigate the raw impact of label noise without using any corrective procedures. However, there is a wealth of research on strategies that mitigate the effects of noisy labels. For completeness and to guide future work, we discuss several such strategies:

- **Robust Loss Functions** : Alternatives to standard cross-entropy, such as Mean Absolute Error (MAE), Generalized Cross-Entropy (GCE), or symmetric cross-entropy (SCE), have demonstrated **resilience to label errors** ([34], [12]). Substituting these in place of the standard loss can reduce the undue influence of mislabeled samples [12]. Although not used here, our future work, particularly with deeper architectures, will adopt such robust losses to mitigate performance degradation.

- **Confident Learning and Data Cleansing:** Confident learning [17] attempts to flag potentially incorrect labels by examining agreement patterns or model predictions. Then, they either relabel or discard questionable cases. We intentionally left noisy labels intact to see their full effect, but real-world practice would likely benefit from auditing and filtering mislabeled samples.

- **Label Smoothing:** A popular technique in neural network training, label smooth- ing replaces hard labels with a softened distribution (e.g., 0.9 for the true class, spread among other classes). This helps when labels may be incorrect, because it prevents the model from overfitting to any single (potentially wrong) label. Although it generally ap- plies to models that train on probabilistic label vectors (e.g., deep nets) and not standard classifiers, label smoothing is a practical and and lightweight way to combat label noise. [35]
- **Semi-Supervised and Re-Labeling Approaches:** Semi-supervised learning can it- eratively refine noisy labels when only a subset of labels are trusted by using the model's confident predictions as pseudo-labels. By treating suspicious samples as "unlabeled" and applying methods such as MixMatch or co-teaching, we can iteratively correct or refine noisy labels. Such methods effectively learn from noisy data by alternating between iden- tifying which labels might be wrong and updating the model, eventually converging to a cleaner label set ([36], [37], [12]). While our scope did not involve unlabeled data or splitting out a trusted subset, semi-supervised strategies can be highly effective when partial label noise is expected, and additional resources allow for iterative data refinement. Noise mitigation methods range from robust losses to fully automated label correction, and each can complement or be combined with standard preprocessing. Our goal was to establish a baseline on unaltered noisy datasets, but our future research will explore these strategies to quantify how effectively they reduce the performance gap caused by mislabeled data.

4.7. Implications Beyond the Healthcare Context

The observed trends—such as simpler models outperforming complex ones under label noise

and class imbalance exacerbating sensitivity degradation—align with findings in non-medical domains. This suggests broader applicability to tasks like fraud detection or equipment fault diagnosis, where label noise and imbalance are common [15].

- Moreover, scaling to larger datasets, such as multi-hospital electronic health record (EHR) repositories or massive image and text corpora, does not necessarily eliminate the under- lying trade-off between model complexity and noise robustness. While distributed train- ing frameworks make it feasible to train large, over-parameterized models on extensive datasets, recent theoretical work by Priebe et al [38] suggests that deep neural networks can tolerate high levels of symmetric label noise, approaching 100% noise under certain asymptotic conditions, essentially achieving Bayes-optimal performance as long as the noise rate is below 100% for each class. In practical terms, this means that if one has a massive dataset (and a model with appropriate capacity and regularization), the model might still learn the correct concept despite many wrong labels, simply because the sheer volume of data offers enough redundancy [38]. For instance, convolutional neural networks have been shown to tolerate 40–50% symmetric noise when paired with careful training strategies such as early stopping ([11],[35]). In contrast, our classical models exhibited steady performance degradation even under moderate noise (10–30%), underscoring that noise impacts vary with model complexity.

- Scalability presents practical challenges. Manual label auditing, while viable for datasets of a few thousand instances, becomes impractical when dealing with hundreds of thousands or millions of samples. In such cases, automated methods such as confident learning [17] are crucial for identifying and mitigating label errors. Benchmark dataset surveys in computer vision, natural language processing, and audio domains report average test set label errors of about 3-4%, which may alter state-of-the-art rankings [17]. Thus, despite the potential advantages of having large volumes of data, high label quality remains essential.

- These considerations extend to various domains where labels often come from noisy prox- ies—finance (e.g., credit risk or fraud detection), advertising (e.g., click or engagement metrics), and other high-throughput industrial settings (e.g., fault diagnosis). Synthetic noise injection, as employed in our study, can help systematically evaluate how sensi- tive different algorithms are to label quality. Across such tasks, ensemble methods and carefully regularized neural networks tend to show relatively greater resilience to noise, although no model is completely immune to its effects.

- Furthermore, computational constraints influence model selection. Scalability issues may rule out of traditionally used classifiers (e.g., SVMs or KNN) in favor of models that support stochastic optimization (e.g., logistic regression or deep networks), especially when combined with noise-robust strategies. Therefore, while our core observation—that label noise impairs performance and that model tolerance differs—is expected to hold across domains, successful deployment at scale will require integrating domain-specific noise mitigation and efficient computational techniques.

.

5. CONCLUSION

This study systematically examined how labeling errors impact the performance of several machine learning models across five medical datasets. Our results reveal that w hile all

models experience performance degradation under noise, the effect on sensitivity is especially pro- nounced in imbalanced settings. Our results show that label noise can substantially impair model accuracy and, more importantly, the ability to correctly identify positive cases (sensitiv- ity) in medical predictions. Even at noise levels as low as 10–15%, we noted the beginning of significant drops in performance for more complex models. Simpler and well-regularized models like logistic regression and SVM demonstrated relatively strong robustness, degrading gracefully as labeling errors increased [13]. In several datasets, these models outperformed more flexible learners once noise passed a critical threshold (around 20% - 25% in many cases). Ensemble methods (random forests) improved over single decision trees in noise tolerance, but still even- tually succumbed to high noise. The use of random oversampling ensured that minority-class performance was improved in baseline conditions; however, it did not immunize the models against noise – in fact, it sometimes amplified the effect of label errors on sensitivity by repli- cating those errors. This highlights a trade-off: oversampling addresses imbalance but must be applied carefully when labels are noisy.

From a practical standpoint, our study suggests that data curation and preprocessing are paramount in medical machine learning. Efforts to reduce label noise – through better label- ing processes, adjudication of disagreements, or algorithmic noise detection – could pay large dividends in model reliability ([5], [16]). When some noise is inevitable, choosing learning al- gorithms that are inherently noise-resistant or incorporating robust training techniques (e.g., loss functions that cap the influence of outliers or training with label noise simulation as data augmentation) may help. Additionally, monitoring metrics like sensitivity and specificity across noise levels provides insight into how a model's errors shift; for instance, noticing a steep sensi- tivity decline might indicate the model has started favoring the majority class due to noise. In imbalanced medical problems (like stroke), maintaining sensitivity is often clinically crucial, so one might accept some loss in specificity to keep sensitivity high – techniques such as adjusting decision thresholds or using cost-sensitive learning could be explored as noise increases. Our comprehensive evaluation reinforces that label noise is a critical factor that can erode the perfor- mance of predictive models in healthcare. We demonstrated clear patterns of which models hold up better and how performance metrics cross over as noise grows. By highlighting these thresh- old effects, we hope this work informs practitioners about the robustness limits of standard ML models. The findings encourage a few concrete next steps. Future research should examine advanced noise-handling methods in the context of medical data – for example, applying noise- robust neural networks or integrating human feedback to correct noisy labels, and comparing those approaches with the baseline degradations we reported. Another direction is to study asymmetric noise reflective of real clinical errors (such as systematically missing diagnoses) and test whether our observed model ranking still holds ([15], [6]). Moreover, developing hybrid solutions that tackle both imbalance and noise jointly (perhaps an improved oversampling that avoids duplicating suspected noisy examples) would be highly valuable. Ultimately, addressing label noise will enhance the trustworthiness and safety of machine learning systems in medicine. By understanding how much noise a given model can handle before it "breaks," stakeholders can make informed decisions on data quality requirements and model selection for deployment. Our work provides a step in that direction, offering a benchmark for noise impact in several representative medical classification tasks and underscoring the need for robust algorithms that can perform reliably amidst the imperfections of real-world data.

6. DECLARATIONS

Conflict of Interest: We declare that there are no conflicts of interest regarding the publication of this paper.

Author Contributions: All authors contributed equally to the effort.

Funding: This research was conducted without any external funding. All aspects of the study, including design, data collection, analysis, and interpretation, were carried out using the resources available within the authors' institution.

Data Availability (including Appendices): All the relevant data, Python code for analysis, detailed annual tables and graphs are available via:

https://github.com/mlprojectbu2025/Impact-of-Noise-on-medical-dataset

REFERENCES

[1] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical image analysis, 65:101759, 2020.

[2] Hongyang Jiang, Mengdi Gao, Yan Hu, Qiushi Ren, Zhaoheng Xie, and Jiang Liu. Label- noise-tolerant medical image classification via self-attention and self-supervised learning. arXiv preprint arXiv:2306.09718, 2023.

[3] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. Journal of artificial intelligence research, 11:131–167, 1999.

[4] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. Artificial intelligence review, 22:177–210, 2004.

[5] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learn- ing with noisy labels. Advances in neural information processing systems, 26, 2013.

[6] David F Nettleton, Albert Orriols-Puig, and Albert Fornells. A study of the effect of differ- ent types of noise on the precision of supervised learning techniques. Artificial intelligence review, 33:275–306, 2010.

[7] Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, Claire Marais Sicre, and Gérard Dedieu. Effect of training class label noise on classification performances for land cover mapping with satellite image time series. Remote Sensing, 9(2):173, 2017.

[8] Bidur Khanal, SM Kamrul Hasan, Bishesh Khanal, and Cristian A Linte. Investigating the impact of class-dependent label noise in medical image classification. In Proceedings of SPIE–the International Society for Optical Engineering, volume 12464, page 1246437, 2023.

[9] Richard W Issitt, Mario Cortina-Borja, William Bryant, Stuart Bowyer, Andrew M Taylor, Neil Sebire, and Stuart A Bowyer. Classification performance of neural networks versus logistic regression models: evidence from healthcare practice. Cureus, 14(2), 2022.

[10] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In International conference on learning representations, 2017.

[11] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694, 2017.

[12] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems, 34(11):8135–8153, 2022.

[13] Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A path to simpler models starts with noise. Advances in neural information processing systems, 36:3362–3401, 2023.

[14] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1944–1952, 2017.

[15] Benoit Frenay and Michel Verleysen. Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems, 25(5):845–869, 2014.
[16] Jakramate Bootkrajang and Ata Kabán. Label-noise robust logistic regression and its applications. In Joint European conference on machine learning and knowledge discovery in databases, pages 143–158. Springer, 2012.
[17] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research, 70:1373–1411, 2021.
[18] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems, 31, 2018.
[19] Jialin Shi, Kailai Zhang, Chenyi Guo, Youquan Yang, Yali Xu, and Ji Wu. A survey of label-noise deep learning for medical image analysis. Medical Image Analysis, 95:103166, 2024.
[20] `Yang Zhao, Zoie Shui-Yee Wong, and Kwok Leung Tsui. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. Journal of healthcare engineering, 2018(1):6275435, 2018.
[21] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021), pages 49–55, 2020.
[22] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. Scientific data, 10(1):1, 2023.
[23] Yixin Wang, Jan GM Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, et al. Gene- expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet, 365(9460):671–679, 2005.
[24] Stroke prediction dataset. https://www.kaggle.com/datasets/fedesoriano/stroke- prediction-dataset.
[25] D Dua and C Graff. Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california, school of information and computer science. IEEE transactions on pattern analysis and machine intelligence, 1(1):1–29, 2019.
[26] Pima indians diabetes dataset. https://www.kaggle.com/datasets/uciml/pima-indians- diabetes-database.
[27] Mengfei Xi, Jie Li, Zhilin He, Minmin Yu, and Fen Qin. Nrn-rsseg: A deep neural network model for combating label noise in semantic segmentation of remote sensing images. Remote Sensing, 15(1):108, 2022.
[28] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5051–5059, 2019.
[29] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost- sensitive learning with noisy labels. Journal of Machine Learning Research, 18(155):1–33, 2018.
[30] Reinier H Stribos. The impact of data noise on a naive bayes classifier. B.S. thesis, University of Twente, 2021.
[31] Marc Ghanem, Abdul Karim Ghaith, Victor Gabriel El-Hajj, Archis Bhandarkar, Andrea de Giorgio, Adrian Elmi-Terander, and Mohamad Bydon. Limitations in evaluating ma- chine learning models for imbalanced binary outcome classification in spine surgery: a systematic review. Brain Sciences, 13(12):1723, 2023.
[32] Mélanie Bernhardt, Daniel C Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C Tez- can, Miguel Monteiro, Shruthi Bannur, Matthew P Lungren, Aditya Nori, Ben Glocker, et al. Active label cleaning for improved dataset quality under resource constraints. Nature communications, 13(1):1161, 2022.
[33] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.
[34] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under la- bel noise for deep neural networks. In Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.

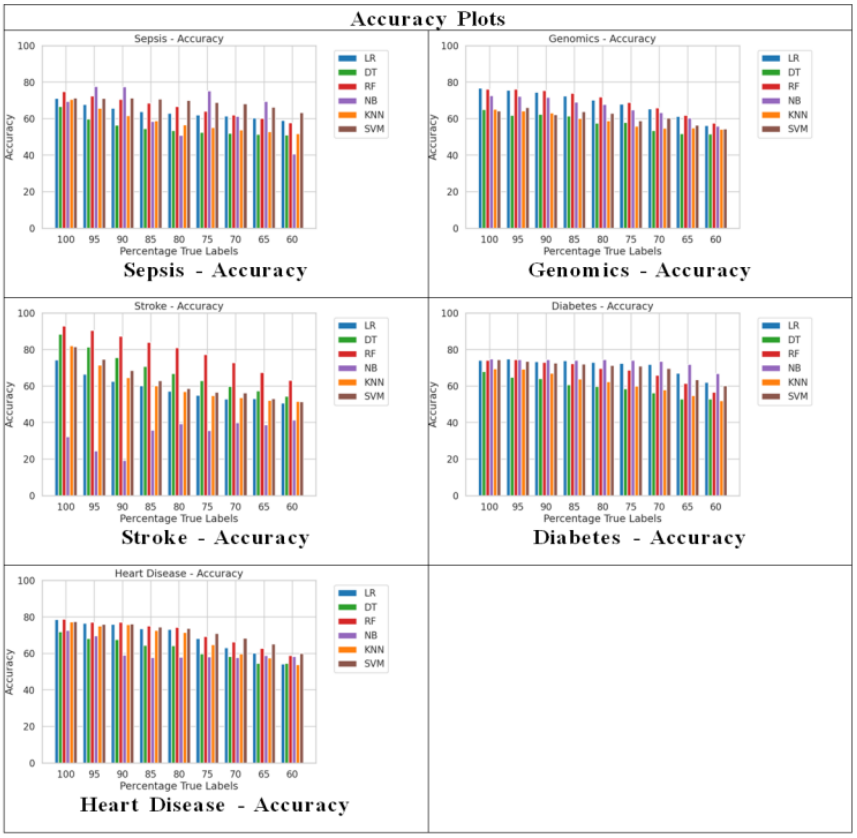
[35] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In International Conference on Machine Learning, pages 6448–6458. PMLR, 2020.

[36] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems, 32, 2019.

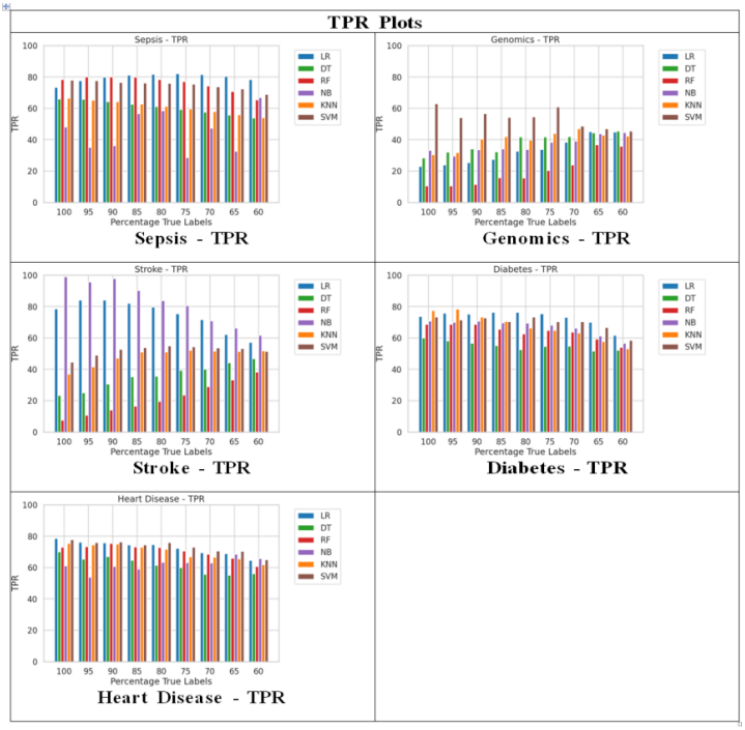
[37] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Robust curriculum learning: from clean label detection to noisy label self-correction. In International Conference on Learning Representations, 2021.

[38] Carey E Priebe, Ningyuan Huang, Soledad Villar, Cong Mu, and Li Chen. Deep learning is provably robust to symmetric label noise. arXiv preprint arXiv:2210.15083, 2022.

APPENDIX A: ACCURACY COMPARISONS



APPENDIX B: TPR COMPARISONS



APPENDIX C: TNR COMPARISONS