Applying Differential Privacy to Training Data

Keaton Banik
Department of Computer Science
University of Western Ontario
Kbanik3@uwo.ca

Darwin Liao
Department of Computer Science
University of Western Ontario
dliao7@uwo.ca

***Abstract*—The growing use of machine learning (ML) in sectors like healthcare and finance has concerns over data privacy, especially with sensitive data governed by regulations like PIPEDA and GDPR. ML models risk leaking private information from reconstruction attacks and model inversion attacks, potentially leading to misuse by malicious actors or unethical practices. Differential Privacy (DP), introduced by Dwork et al. in 2006, provides a mathematical framework to safeguard privacy by limiting the extraction of sensitive data from models. By integrating DP, organizations can leverage ML for insights while adhering to privacy standards and protecting individual confidentiality, however, in doing so, ML models will suffer in performance due to the nature of noise in training data. In this study we attempt to apply different DP mechanisms at differing privacy budgets to a sample synthetic dataset, and then gauge the performance of a Convolutional neural network (CNN) when trained on the DP datasets, and see what factors could be important when using a DP dataset in training as well as some factors that may affect the level of privacy.**

I. Introduction

Recently, the popularity of machine learning and AI has significantly increased among many different sectors such as, but not limited to, health and finance. These different sectors generally work with big data and would benefit greatly from powerful data processing mechanisms, allowing companies to use its data to generate meaningful conclusions such as targeting potential advertisements to recommended treatments. However, the data that is being used and collected is oftentimes protected by both national regulations and company regulations such as PIPEDA and GDPR, and without proper protection in place that conform to the regulations in all steps of the data processing, there runs a risk of extremely sensitive data being leaked.

This issue of data privacy exists in current machine learning models, where sensitive data can be extracted from the model unless there are implementations in place, and could lead to detrimental use of the extracted information. Some examples include situations where insurance companies may raise the cost for certain customers when they obtain the information that these customers have prevailing medical ailments, or bad actors obtaining sensitive information and using this information to track and blackmail individuals. One such example of this incident was done in [1], where a significant portion of private data in the 2010 US census was reconstructed from deidentified data.

These concerns and incidents have led to the question of how privacy can be defined, and in 2006, Dwork et al introduced a concept called Differential Privacy (DP) [2] which is often considered the gold standard of privacy definitions, and provides a mathematically precise and quantifiable notion of privacy. The concept of DP is built upon the idea of the Fundamental Theory of Information Recovery which states that overly accurate answers to too many questions will destroy privacy in a spectacular way [3]. This idea is directly relevant to machine learning, where models are trained on vast datasets, often containing sensitive individual information. Machine learning models aim to generalize patterns and make predictions based on input data without memorizing specific details about individual entries. However, without safeguards like DP, models may inadvertently ”leak” sensitive information during inference or adversarial attacks.

II. Previous Work

A. Data Privacy

The topic of data privacy has been a long-studied field for the past few decades, and there exists a large amount of literature involved. However, there are a few outstanding methods that have been influential in the development of modern definitions and methods in data privacy. In 1998, an early method of data privacy was proposed by Sweeney known as k-anonymity [1]. This method hides data and makes individuals indistinguishable from at least k-1 other people, but is extremely costly to perform on larger data sets and high values of k, making it not feasible for big data applications and machine learning. In 2006, Dwork et al proposed DP [2], with the mathematically quantifiable definition of privacy as shown (1).

$$\ln\left(\frac{\Pr[\mathcal{M}(x) \in S]}{\Pr[\mathcal{M}(y) \in S]}\right) \leq \epsilon \tag{1}$$

Where Pr is the probability, M is a randomization mechanism applied to the datasets x and y, where x and y represent two neighboring datasets that differ by only one individual’s data. Differential privacy ensures that the mechanism’s output is similar whether or not a particular individual’s data is included. S is the subset of possible outputs of the mechanism M, and Epsilon is the privacy guarantee. A smaller Epsilon represents a stronger privacy guarantee because the outputs of x and y can differ less.

This base definition of DP has been expanded upon in the past decade, and many variations of DP have been proposed, all of which aim to target specific weaknesses in the original definition. In the same 2006 paper, Dwork et al also introduced (ϵ, δ)-Differential Privacy, where a slack variable δ is included into the definition to address the issue of the original definition being too strict, known as approximate DP. In 2016, Bun et al introduced another variation of DP known as Zero-Concentrated DP, where instead of bounding the probability ratio between outputs for neighboring datasets (as in traditional DP), Zero-CDP focuses on the Rényi divergence between the output distributions as shown in (2)

$$D_\alpha(M(D) \parallel M(D')) \leq \rho \cdot \alpha, \tag{2}$$

Where D_α is the the Rényi divergence of order α between two probability distributions P(M(D)) and Q(M(D')). ρ is the privacy loss parameter, which controls the level of privacy guarantee. In 2020, Dong et al proposed a different method of Gaussian Differential Privacy (GDP), which introduces a single parameter, μ, to characterize privacy derived from the Rényi divergence between the probability distributions of the mechanism outputs under neighboring datasets and also quantifies privacy loss in terms of the hypothesis testing framework. This method results in the tightest possible privacy bound for the Gaussian mechanism. [5]

B. Mechanisms

In the previously referenced definitions of DP, there exists M, a randomization mechanism. There are many randomization mechanisms that are used in DP, the Laplace mechanism, the Gaussian mechanism, and exponential mechanism are all randomization mechanisms that have been used for different data types situations where one mechanism would not be suitable but another would. When used in the base definition of DP, they have all been shown to be ϵ-differentially private.[5]

C. Differential Privacy in Machine Learning

Different privacy has been shown to be useful in machine learning situations due to the fact that training complex models tend to use a large amount of data, which can be exploited by users and compromise the privacy of the data. There have been various studies showing how reconstruction attacks and model inversion attacks can infer sensitive information from models [7][8], and shows the importance of having private learning. There is a significant body of work on this topic, which can generally be classified into three main approaches: pre-processing, in-processing, and post-processing. Pre-processing includes methods that apply DP to the data before the model is trained on it, and post-processing applies DP to the final results, as shown by Chaudhuri et al. in [9]. In-processing covers methods that apply differential privacy to the individual steps of the learning process. One such method is adding noise to the gradient descent process as shown in [9].

III. Experimental Setup

The goal is to analyse the effects that preprocessing databases using differential privacy have on a machine learning model, and how different pre-processing methods compare to each other in a real world setting. To accomplish this, a sufficiently complex dataset must be used to both simulate a big data scenario, and to ensure that there would be enough data to apply differential privacy without the effects being too drastic for the model to handle.

The dataset selected was the Synthea Covid-19 dataset which can be found here: [10]. This is a synthetic dataset that simulates data obtained from patients and is free from any data, privacy, and security restrictions. To facilitate the process of designing a simple model, not all portions of the dataset were used in this project, as some were deemed either too complex or unnecessary to include for the scope of this experiment, such as payment history and methods, and location of hospital.

The dataset used in this study was derived from Synthea, an open-source patient population simulator designed to generate synthetic but realistic healthcare data. Synthea produces datasets that mimic real-world populations and their interactions with healthcare systems while ensuring that no actual patient data is used, thus maintaining privacy and security. Synthetic data allows for the evaluation of differential privacy techniques in machine learning models without risking the exposure of sensitive real-world data. Since the primary aim is to examine the effectiveness of privacy-preserving techniques rather than derive clinical insights, synthetic data serves as an ethical and practical alternative. The generated data is comprehensive, covering patient demographics, medical conditions, care plans, medications, observations, allergies, immunizations, and other relevant healthcare attributes. The generated data is comprehensive, covering patient demographics, medical conditions, care plans, medications, observations, allergies, immunizations, and other relevant healthcare attributes. The COVID-19-specific module of Synthea simulates the spread, treatment, and outcomes of COVID-19 within a synthetic population. Out of all the available data, we included allergies, observations, conditions, immunizations, medications, and patients due to their direct medical impact on COVID-19 contraction and mortality, as well as patient information such as name, age, and city. The detailed documentation and methodology behind Synthea's data generation can be accessed through the project's official documentation [10]. Additional insights into the clinical validity and structure of the data are discussed in a related study [11].

After the initial cleaning, the randomization mechanisms were applied to each column depending on the context of the column and the type of data retained. Two main mechanisms were applied to the dataset's columns, each dependent on the type of data, the Laplace and Exponential Mechanisms. The Laplace mechanism is simply a Laplace distribution added onto the values of a dataset, given by:

$$f(x; \mu, \lambda) = \frac{1}{2\lambda} \exp\left(-\frac{|x - \mu|}{\lambda}\right) \tag{3}$$

Where μ in this case would always be 0 because the distribution would be noise added onto the data point, and lambda would be the scale of the function. This lambda is calculated through the sensitivity, which was set as 1 and the privacy factor ϵ shown in (1). For each of the datasets that used a Laplace mechanism, ϵ was set at 0.5, 1, and 3 for three separate levels of privacy. The Exponential mechanism was used in place of the Laplace mechanism in cases where data was categorical and it did not make sense for a continuous distribution to be applied. Information such as cause of death, positive or negative testing of certain conditions, and recreational drug use were examples of categories in which the exponential mechanism was used. The mechanism is defined by:

$$\Pr[o] \propto \exp\left(\frac{\epsilon \cdot u(D,o)}{2\Delta u}\right) \tag{4}$$

Where $o \in O$ and O is a set of all possible outputs. ϵ is the privacy factor, u is the utility function which sets a utility value on each value of o on dataset D, and $\delta\mu$ is the sensitivity of the utility function, which was also set at 1. The utility function is a function that determines how desirable an output is, where a higher value is valued more. In the experiment, the utility function was calculated based on the frequency of each output. However, unlike the Laplace mechanism, the privacy factor ϵ was only set at 0.5 and 1 due to the exponential scaling of the mechanism. datasets.

IV. Results and Analysis

A. Data Distribution Before and After SMOTE

Figure 1 illustrates the class distribution in the dataset before and after applying the Synthetic Minority Oversampling Technique (SMOTE). Before SMOTE, a significant class imbalance was evident, with Class 3 (COVID+Lived) being overrepresented while Class 0 (COVID+Died) was severely underrepresented.

After applying SMOTE, the dataset achieved uniform class distribution, as seen in the right panel of Figure 1. By generating synthetic samples for the minority classes, SMOTE effectively balanced the dataset, ensuring fairer training and better generalization across all classes.

For datasets utilizing differential privacy random noise, Figure 2 (ϵ = 0.5), Figure 3 (ϵ = 1), and Figure 4 (ϵ = 3) demonstrate the class distributions before and after applying SMOTE. Across all three noise levels, similar trends were observed. Before SMOTE, Class 3 remained overrepresented, while Class 0 and Class 1 exhibited significant underrepresentation. Post-SMOTE balancing effectively harmonized the class distributions in all cases, as reflected by the equalized frequencies in the right panels of the respective figures.

Notably, the magnitude of imbalance differed slightly between the base model and the differential privacy datasets. For instance, at ϵ = 0.5, Class 3 exhibited a slightly higher initial overrepresentation compared to the base dataset, while ϵ = 3 showed a more pronounced imbalance in Class 1. Despite these initial differences, the application of SMOTE consistently corrected the imbalance, demonstrating its robustness across varying data distributions.

B. Training and Validation Performance

The base model was trained over 23 epochs, as shown in Table 1, Figures 5 and Figure 6. Training accuracy steadily improved, beginning at 81.33% in the first epoch and reaching 94.35% by the 23rd epoch. Validation accuracy followed a similar trend, starting at 93.17% and stabilizing at approximately 96.2%. Training loss showed a substantial reduction, declining from 4.16 to 0.5366, while validation loss decreased from 1.6297 to 0.454, indicating successful learning and reduced overfitting. Figure 5 illustrates the convergence of training and validation accuracy over the epochs, highlighting consistent improvement. Figure 6 depicts the training and validation loss, showing clear optimization and decreasing error rates across the epochs.

For datasets with differential privacy random noise, Table 1 and Figures 7 through 12 summarize training and validation performance across different levels of noise. At ϵ = 0.5, training accuracy began at 70.28% and reached 88.63% by epoch 39, while validation accuracy improved from 73.73% to 94.44%. Training loss reduced from 4.6589 to 0.9436, and validation loss decreased from 1.9908 to 0.8265, as illustrated in Figures 7 and 8. For ϵ = 1, training accuracy increased from 73.54% to 91.78% by epoch 24, with validation accuracy rising from 76.18% to 97.59%. Training loss dropped from 4.6345 to 1.1309, while validation loss declined from 2.2406 to 1.0549, as shown in Figures 9 and 10. Finally, at ϵ = 3, training accuracy improved from 69.96% to 90.91% over 29 epochs, with validation accuracy rising from 83.83% to 96.63%. Training loss decreased from 4.2683 to 0.9449, and validation loss declined from 1.6709 to 0.7774, as depicted in Figures 11 and 12.

Table 1 shows that although the base model achieved faster convergence in training accuracy compared to datasets with differential privacy noise, the training and validation losses were significantly higher in the randomized datasets. For example, at ϵ = 0.5, the validation loss remained around 0.8265, which is relatively high compared to the base model's final validation loss of 0.454. These findings suggest that the addition of differential privacy noise adversely impacts the model's ability to minimize error, resulting in reduced overall performance.

When training the model on differentially private datasets, it was shown that as ϵ was introduced, the amount of epochs required greatly increased, from 23 to 40 epochs. As ϵ increased (decreasing privacy), the amount of epochs required quickly returned to the original amount, and the model was able to adjust to the low amount of additional noise introduced in the dataset. This follows the expected results for features that use a laplace mechanism due to the fact that the value of ϵ is inversely proportional to the noise generated, as shown in (3).

C. Test Performance

The performance of the base model was evaluated on both the original and SMOTE datasets. On the original dataset, as summarized in Table 2, the model achieved a test loss of 0.4917 and a test accuracy of 96.68%. In comparison, on the SMOTE dataset, the model exhibited improved performance with a test loss of 0.4424 and a test accuracy of 96.77%.

For models trained with differential privacy random noise, Table 2 highlights a decline in test performance as the level of noise increases. For the SMOTE dataset, the model trained with ϵ = 0.5 achieved a test loss of 0.7755 and an accuracy of 95.28%. At ϵ = 1, the test loss further increased to 0.9079, with a slight improvement in accuracy at 97.57%. The model trained with ϵ = 3 exhibited a test loss of 0.7579 and an accuracy of 96.50%. Similarly, for the original dataset, the test loss and accuracy were negatively impacted by the addition of noise. The base model achieved a loss of 0.4917 and an accuracy of 96.68%, whereas models trained with ϵ = 0.5, ϵ = 1, and ϵ = 3 exhibited test losses of 0.8747, 0.9753, and 0.8248, respectively, with corresponding accuracies of 89.15%, 94.17%, and 93.20%.

These results emphasize that the randomized models generally performed worse compared to the base model, as evidenced by the significantly higher test losses across both the SMOTE and original datasets. While the test accuracies for higher values of e approached the base model's performance, the elevated loss values indicate suboptimal model optimization and increased error rates.

D. Classification Report

The classification report on the original dataset (Table 3) reveals the model's detailed performance metrics across all classes. Precision ranged from 92% to 99%, recall from 82% to 99%, and F1-scores from 87% to 98%. Specifically, Class 0 (COVID+Died) achieved a precision of 94%, recall of 97%, and an F1-score of 95%. For Class 1 (NoCOVID+Died), the precision was 99%, recall 96%, and F1-score 98%. Class 2 (NoCOVID+Lived) had a precision of 92%, recall of 82%, and an F1-score of 87%. Lastly, Class 3 (COVID+Lived) demonstrated precision, recall, and F1-scores of 97%, 99%, and 98%, respectively. Overall, the model achieved a test accuracy of 97%, with macro-averaged precision, recall, and F1-scores of 95%, 94%, and 94%, respectively.

On the SMOTE dataset, the classification report (Table 4) demonstrates consistent improvement in recall and F1-scores across all classes. Class 0 (COVID+Died) achieved perfect metrics with a precision, recall, and F1-score of 100%. Class 1 (NoCOVID+Died) achieved a precision of 99%, recall 97%, and F1-score 98%. Class 2 (NoCOVID+Lived) saw significant improvement with a precision of 97%, recall of 91%, and an F1-score of 94%. Class 3 (COVID+Lived) also improved with a precision of 92%, recall of 99%, and an F1-score of 95%. The overall test accuracy on the SMOTE dataset was 97%, with macro-averaged precision, recall, and F1-scores of 97%.

However, the prevalence of perfect scores (e.g., precision and recall of 1) in SMOTE-generated data reveals potential issues with the SMOTE method, as it may artificially inflate performance metrics. This observation suggests that while SMOTE improves class balance, it may compromise the evaluation of true model performance by overrepresenting specific patterns.

For models trained with differential privacy random noise on the original dataset, substantial variability in classification metrics was observed, highlighting the impact of noise addition. For instance, at ϵ = 0.5, Class 3 demonstrated a low precision of 43% and an F1-score of 52%, accompanied by a perfect recall of 100%. However, this recall score was based on only two samples, significantly reducing the support from 74 in the base model to 2, which highlights how randomization affects the classification of COVID-related outcomes and mortality. Similar trends were seen across other classes, where the support values dropped, reflecting the challenges posed by privacy-induced randomness.

As the noise level increased, the performance of randomized models gradually improved. At ϵ = 3, macro-averaged precision, recall, and F1-scores increased to 77%, 91%, and 82%, respectively. These results indicate that higher noise levels may mitigate some of the adverse effects of differential privacy while still impacting overall accuracy and class-specific performance metrics.

E. Confusion Matrix Analysis

The confusion matrices for both SMOTE and original datasets provide insights into the predictive performance of the models under different conditions. For the SMOTE dataset, the base model (Figure 13) displayed strong performance, with minimal misclassifications primarily occurring in Class 2 (NoCOVID+Lived), where 139 instances were incorrectly classified as Class 3 (COVID+Lived). Similar trends were observed in the models trained with differential privacy noise. At ϵ = 0.5 (Figure 15), misclassifications increased, particularly in Class 2, while ϵ = 1 (Figure 17) showed improved predictions. At ϵ = 3 (Figure 19), the model achieved better stability, with reduced errors in Class 2 and 3 compared to lower noise levels. However, the addition of noise still affected the differentiation between these classes.

On the original dataset, the base model (Figure 14) exhibited more notable errors due to class imbalance. Class 2 was frequently misclassified as Class 3, and support for Class 0 (COVID+Died) dropped significantly in randomized models. At ϵ = 0.5 (Figure 16), severe misclassifications were observed in Classes 0 and 3, where low support further exacerbated prediction challenges. The trend continued at ϵ = 1 (Figure 18), where moderate improvements were evident, but support values remained disproportionately low. By ϵ = 3 (Figure 20), the model achieved improved stability and accuracy, although misclassifications persisted in underrepresented classes.

The confusion matrices reveal important trends and challenges. SMOTE-based models generally demonstrated superior performance compared to models trained on the original dataset, as the balanced class distributions reduced the negative impact of underrepresented labels. However, SMOTE failed to adequately mimic the interactions inherent in the original patient dataset, leading to less realistic classifications and potentially overfitting to synthetic patterns.

Differential privacy randomization significantly affected classification performance. For instance, in the randomized models, Class 0 (COVID+Died) and Class 3 (COVID+Lived) exhibited drastic drops in support, with support for Class 0 in the original data falling from 74 to just 2 instances under ϵ =

0.5. This resulted in high recall but very low precision and F1-scores for these classes, exposing the limitations of randomization in preserving meaningful patterns related to COVID and mortality. Similar patterns emerged in other classes, where misclassifications increased due to reduced support and data noise.

As the noise level increased, some improvements in stability and accuracy were observed, particularly at $\epsilon = 3$, where better differentiation between classes became apparent compared to lower noise levels. Nevertheless, misclassifications persisted, particularly in underrepresented classes, demonstrating the trade-off between preserving privacy and achieving predictive accuracy. These results highlight the necessity of balancing class distributions while addressing the difficulties introduced by differential privacy noise to enhance model performance without compromising the integrity of COVID and mortality classification.

F. Feature Importance Analysis

The analysis of the top five features contributing to the model's decision-making, as summarized in Table 5, reveals both consistencies and differences between the base and randomized models. For the base model, ”Fibrin D-dimer FEU in Platelet poor plasma” exhibited the highest correlation (0.367), highlighting its critical role in predicting COVID-related outcomes. Other influential features included ”Adenovirus A+B+C+D+E DNA” (0.350) and ”Human metapneumovirus RNA” (0.350), both identified from respiratory specimens. Features such as ”Influenza virus A RNA” and ”Influenza virus B RNA” also showed notable correlations of 0.340 and 0.338, respectively.

In contrast, the randomized models prioritized features differently, with ”SARS-CoV-2 RNA Panel in Respiratory NAA+probe” consistently emerging as the most critical across all noise levels ($\epsilon = 0.5$, $\epsilon = 1$, and $\epsilon = 3$). Its correlation ranged from 0.489 at $\epsilon = 0.5$ to 0.528 at $\epsilon = 1$, highlighting its resilience under privacy-induced randomness. This feature reflects the primary importance of SARS-CoV-2 RNA in understanding the progression and mortality related to COVID, aligning with clinical significance.

Other recurring features in the randomized models included ”Left ventricular Ejection fraction” and various patient-level metrics. For example, at $\epsilon = 0.5$, ”AGE” was identified as a significant factor, with a correlation of 0.279, while at $\epsilon = 1$ and $\epsilon = 3$, clinical observations such as ”QALY” (Quality-Adjusted Life Years) and specific medications (e.g., ”Metoprolol” and ”Furosemide”) gained prominence. These findings suggest that noise addition shifts the model's reliance toward broader clinical indicators and away from granular biomarkers observed in the base model.

A notable divergence is the reduced emphasis on respiratory virus indicators such as ”Influenza RNA” in the randomized models. Instead, medication-related features, such as ”Metoprolol succinate” and ”Chronic congestive heart failure (disorder),” became increasingly important at higher noise levels ($\epsilon = 3$). This shift likely reflects the impact of randomization on feature distribution and the model's attempt to optimize predictions under constrained data utility.

Interestingly, while the base model's features align closely with known COVID-related biomarkers, the randomized models demonstrate a broader reliance on systemic and patient-level features, potentially due to the suppression of precise patterns by noise. This shift shows the trade-offs inherent in differential privacy: as noise levels increase, critical medical insights may be diluted, complicating the interpretability and clinical applicability of the model.

Among the top five features, Adenovirus, Metapneumovirus, and both Influenza features were binary categorical data subjected to an exponential mechanism with ϵ=1, likely significantly contributing to the slight inaccuracy increase compared to the ϵ=3 test runs, as the ability to only have one of two options be present would greatly skew the outcome of the model's predictions when stricter privacy was enforced. The exponential mechanism's sensitivity to binary categorical features amplifies the impact of noise, as the limited range of options (e.g., presence or absence) results in a higher likelihood of misclassification or skewed utility values.

In summary, the feature importance analysis demonstrates that the base model leverages precise biomarkers closely tied to COVID-19 outcomes, while randomized models shift focus toward broader clinical features. These trends reveal the significant effects of privacy constraints on feature selection and the critical need for optimized approaches to maintain both privacy and predictive utility.

V. Discussion

In general, using a clean dataset when training a prediction model generated significantly better results compared to using a dataset with differential privacy noise applied. This outcome aligns with expectations given the trade-offs inherent in differential privacy mechanisms, where noise addition ensures privacy but compromises data utility. This trade-off highlights the core challenge of selecting an appropriate privacy budget (ϵ) that balances performance and privacy.

The experiments demonstrated that applying differential privacy had varying impacts depending on the sensitivity of the affected features. Specifically, features with high importance to the model's decision-making were more adversely affected by stricter privacy budgets, leading to significant reductions in model accuracy. This finding emphasizes the necessity of tailoring privacy mechanisms to account for feature significance, as indiscriminate application of noise can disproportionately degrade model performance.

The study also identified issues related to class imbalance, particularly for cases where patients contracted COVID-19 and died. The imbalance skewed the model's predictions toward more prevalent classes. This effect was compounded by the exponential mechanism's reliance on frequency-dependent utility functions, which further obfuscated minority class outcomes under stricter privacy constraints. The Synthetic Minority Oversampling Technique (SMOTE) effectively addressed these imbalances in the dataset. However, while SMOTE improved class representation and overall performance, it also introduced potential distortions by over representing synthetic patterns, which may not reflect real-world scenarios accurately.

Differential privacy's impact on feature importance revealed a notable shift in the model's reliance on certain features. While the base model prioritized precise biomarkers closely tied to COVID-19 outcomes, the randomized models under differential privacy noise relied more on broader clinical features. This shift indicates that higher noise levels dilute the utility of granular, high-impact features, complicating interpretability and clinical applicability. Moreover, binary categorical features subjected to the exponential mechanism displayed heightened susceptibility to noise, further underscoring the importance of carefully selecting privacy mechanisms based on data type and application context.

VI. Conclusions

This study evaluated the effects of differential privacy techniques on the performance of machine learning models using synthetic healthcare data derived from the Synthea COVID-19 dataset. The findings confirmed that the addition of noise through differential privacy mechanisms adversely impacts model accuracy and requires more epochs for convergence. Despite these challenges, higher ϵ values (lower privacy levels) allowed models to achieve improved accuracy and reduced training loss, suggesting a viable pathway for balancing privacy and performance.

The results underline the importance of thoughtful privacy budget selection and feature prioritization when applying differential privacy. Factors such as class imbalance, feature sensitivity, and data type significantly interact with privacy mechanisms, amplifying their effects on model performance. Techniques like SMOTE can mitigate some of these challenges, but they also introduce their own set of limitations, particularly concerning the realism of synthetic data.

Future research should explore adaptive privacy mechanisms that dynamically adjust noise levels based on feature importance or class representation. Additionally, integrating differential privacy with other methodologies, such as federated learning or hybrid privacy approaches, could enhance both data privacy and model utility. Addressing these areas will be critical for advancing the practical application of differential privacy in sensitive domains like healthcare while ensuring ethical and effective data use.

References

[1] S. Garfinkel, J. M. Abowd, and C. Martindale, ”Understanding Database Reconstruction Attacks on Public Data,” Communications of the ACM, vol. 62, no. 3, pp. 46–53, Mar. 2019, doi: 10.1145/3287287.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, ”Calibrating Noise to Sensitivity in Private Data Analysis,” Journal of Privacy and Confidentiality, vol. 7, no. 3, pp. 17–51, May 2017, doi: 10.29012/jpc.v7i3.405.

[3] [3]C. Dwork and A. Roth, The Algorithmic Foundations of Differential Privacy, Foundations and Trends® in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014, doi: 10.1561/0400000042.

[4] L. Sweeney, ”k-Anonymity: A Model for Protecting Privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002, doi: 10.1142/S0218488502001648.

[5] J. Dong, A. Roth, and W. J. Su, ”Gaussian differential privacy,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 84, 2022.

[6] J. Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 94–103. https://doi.org/10.1109/FOCS.2007.66

[7] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), 3–18. https://doi.org/10.1109/SP.2017.41

[8] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep Learning with Differential Privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16), 308–318. https://doi.org/10.1145/2810103.2813677

[9] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, ”Differentially private empirical risk minimization,” Journal of Machine Learning Research, vol. 12, no. 29, pp. 1069–1109, 2011. [Online]. Available: http://jmlr.org/papers/v12/chaudhuri11a.html.

[10] O. Williams and F. McSherry, ”Probabilistic inference and differential privacy,” in Advances in Neural Information Processing Systems (NeurIPS), J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23. 2010.

[11] MITRE Corporation, ”Synthea: Synthetic Patient Population Simulator.” [Online]. Available: https://synthea.mitre.org/downloads. [Accessed: Dec. 24, 2024].

[12] J. Walonoski, S. Klaus, E. Granger, D. Hall, A. Gregorowicz, G. Neyarapally, A. Watson, and J. Eastman, “Synthea™ Novel coronavirus (COVID-19) model and synthetic data set,” Intelligence-Based Medicine, vol. 1–2, p. 100007, 2020, doi: 10.1016/j.ibmed.2020.100007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666521220300077.

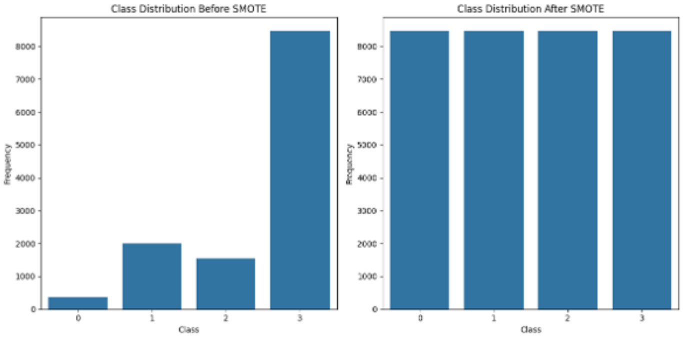


Fig. 1. Class Distribution for the Original Dataset Before and After SMOTE

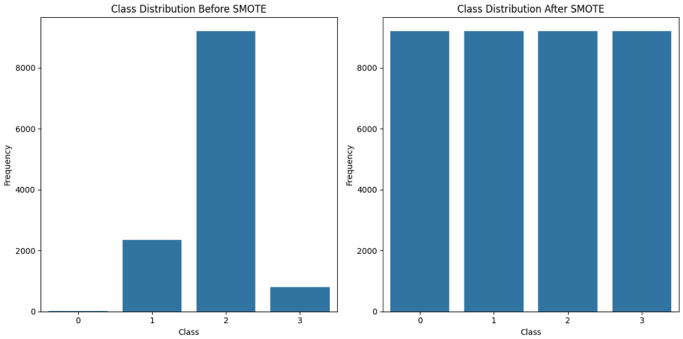


Fig. 2. Class Distribution for the ϵ = 0.5 Dataset Before and After SMOTE

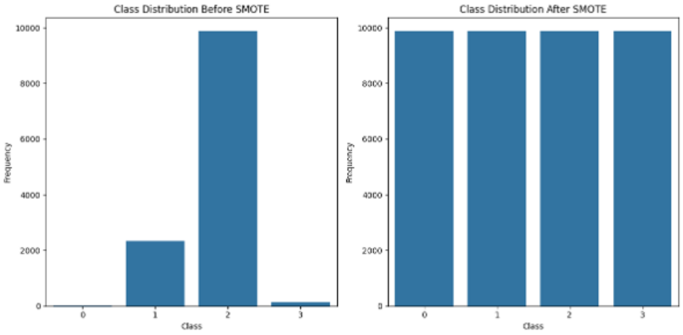


Fig. 3. Class Distribution for the $\epsilon = 1$ Dataset Before and After SMOTE

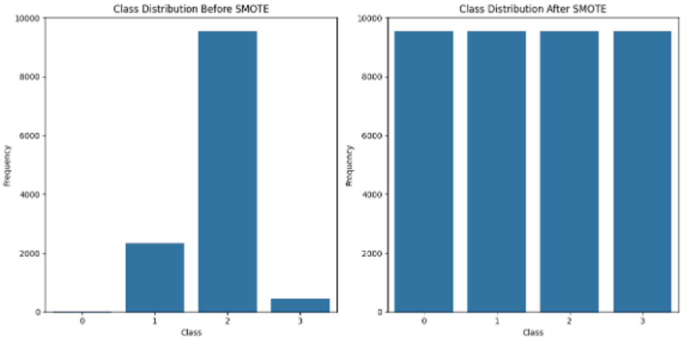


Fig. 4. Class Distribution for the $\epsilon = 3$ Dataset Before and After SMOTE

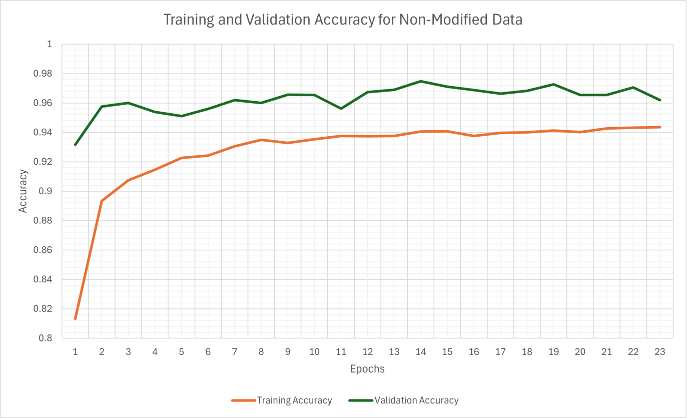


Fig. 5. Training and Validation Accuracy for Original Data

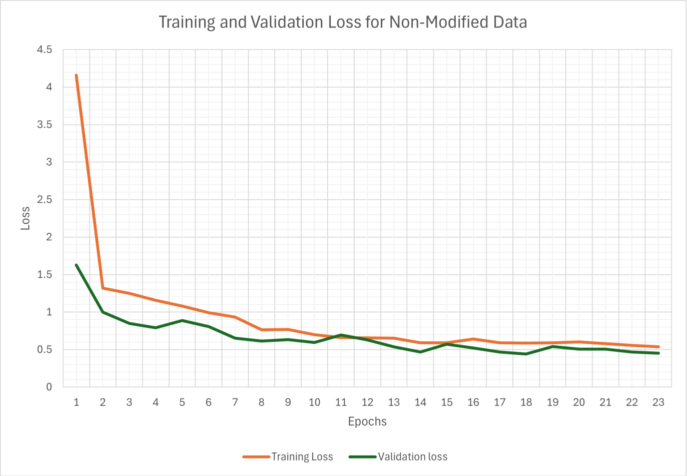


Fig. 6. Training and Validation Loss for Original Data

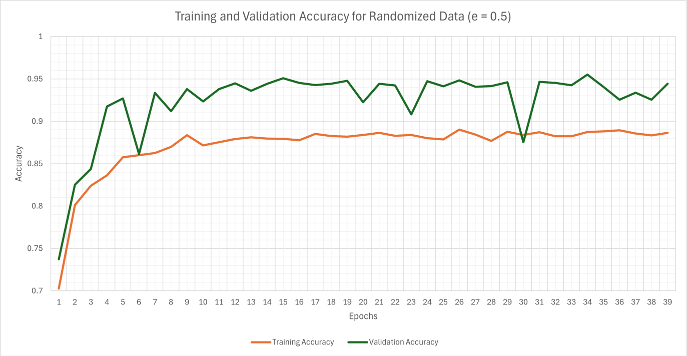


Fig. 7. Training and Validation Accuracy for $\epsilon = 0.5$

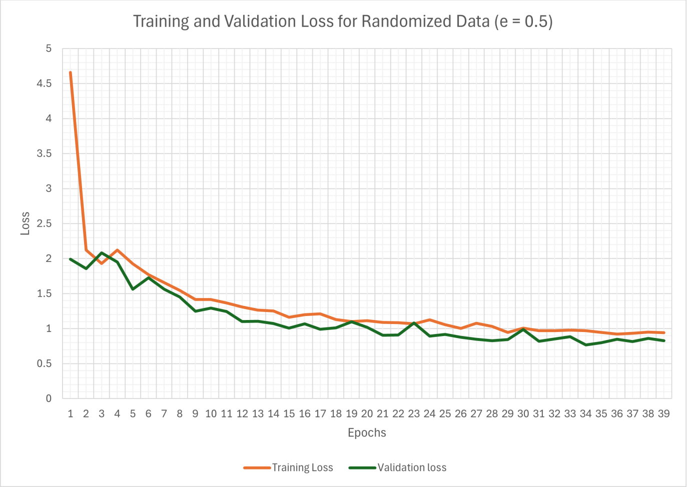


Fig. 8. Training and Validation Loss for $\epsilon = 0.5$

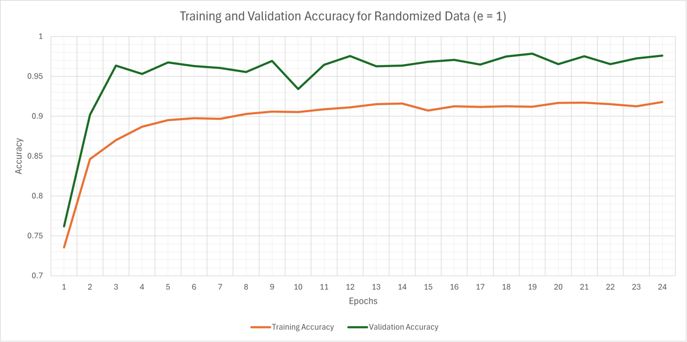


Fig. 9. Training and Validation Accuracy for ϵ = 1

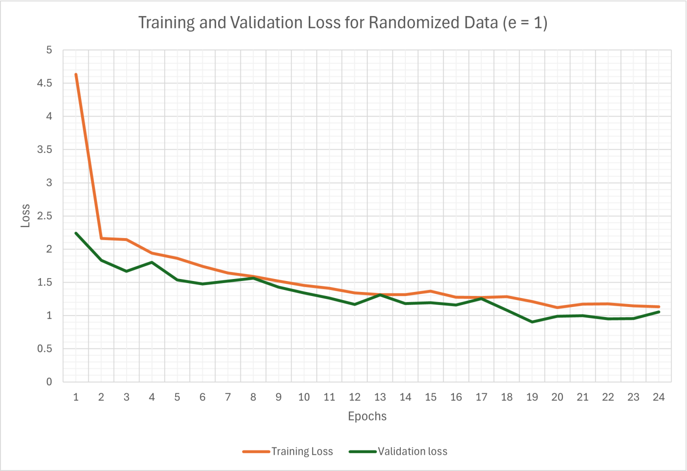


Fig. 10. Training and Validation Loss for ϵ = 1

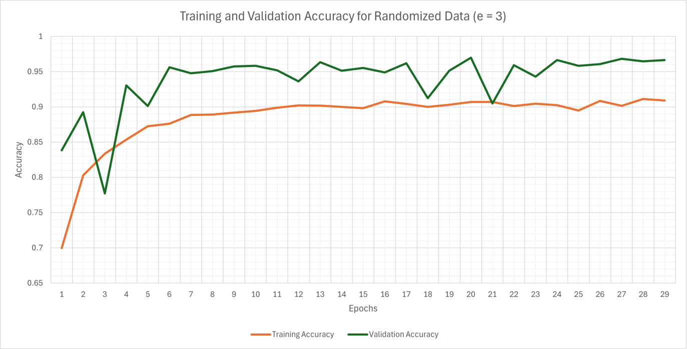


Fig. 11. Training and Validation Accuracy for ϵ = 3

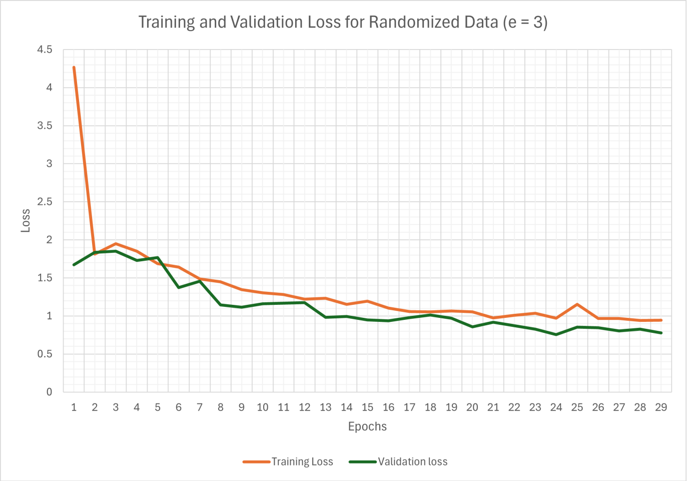


Fig. 12. Training and Validation Loss for ϵ = 3

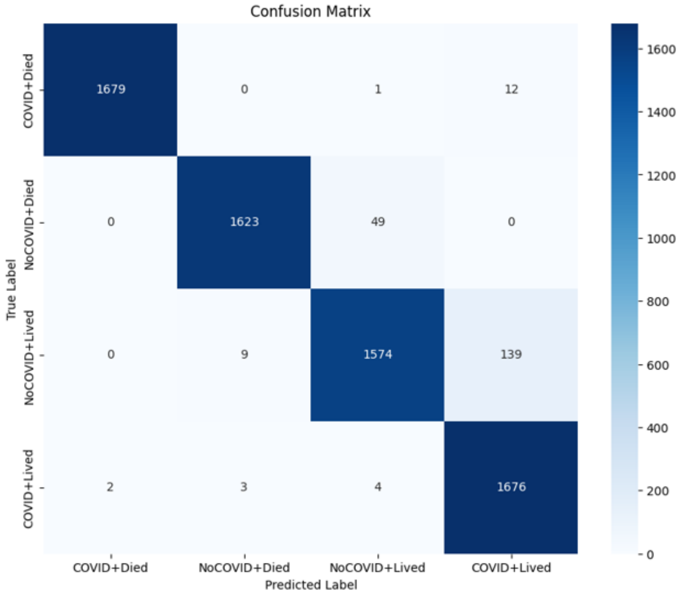


Fig. 13. Confusion Matrix of the Original Model with SMOTE data

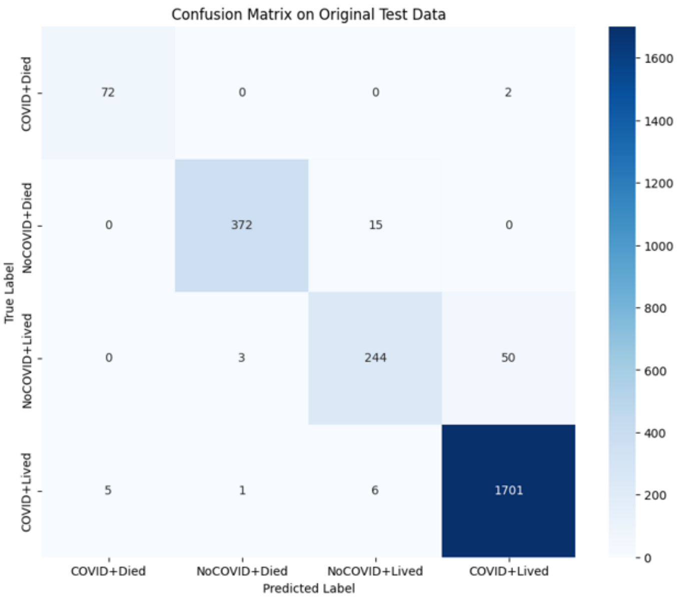


Fig. 14. Confusion Matrix of the Original Model with Original Data

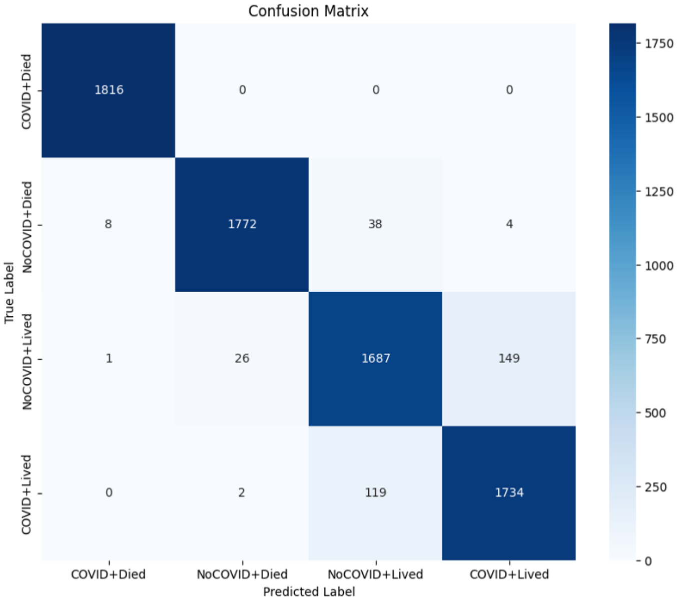


Fig. 15. Confusion Matrix of the ϵ = 0.5 Model with SMOTE data

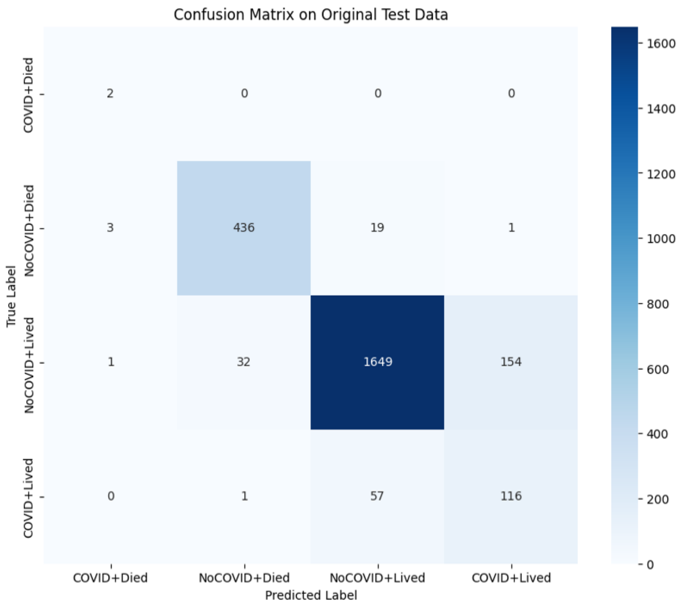


Fig. 16. Confusion Matrix of the ϵ = 0.5 Model with Original Data

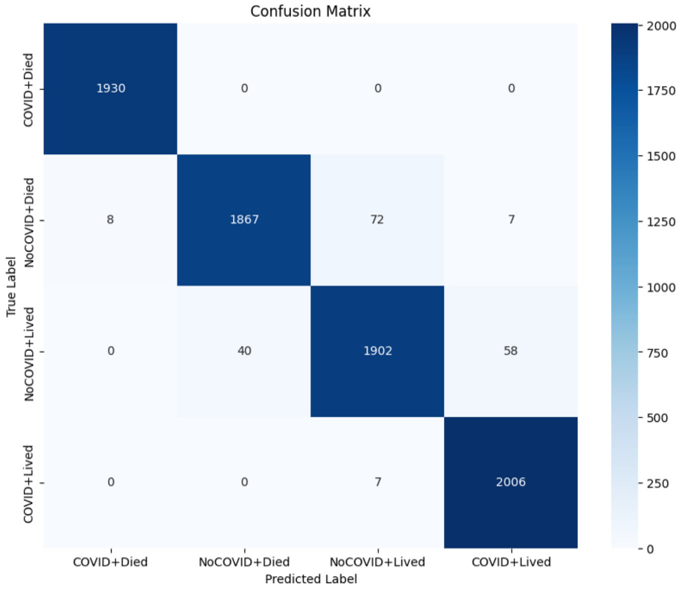
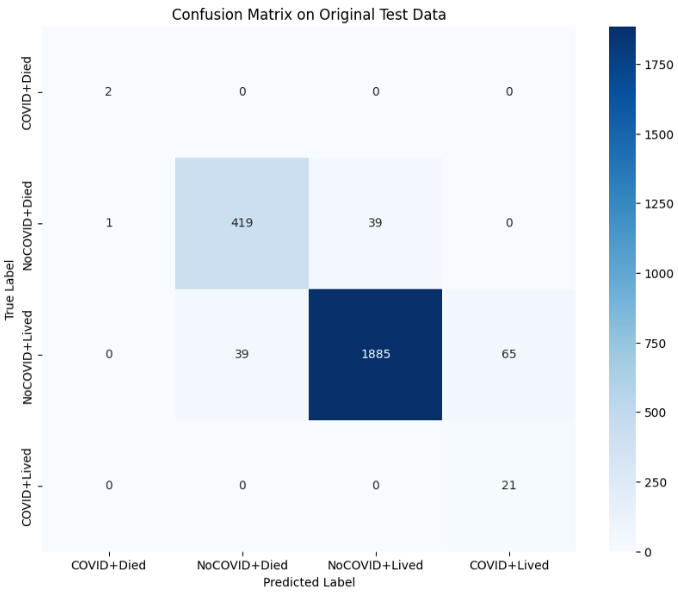
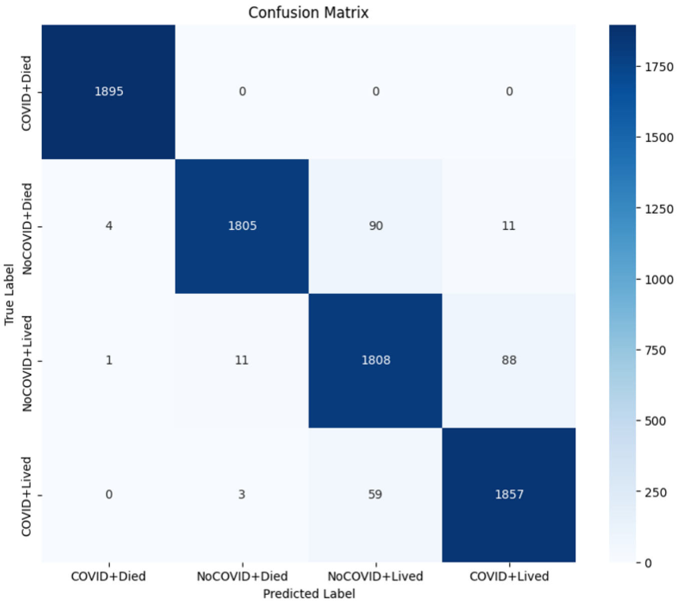


Fig. 17. Confusion Matrix of the ϵ = 1 Model with SMOTE data



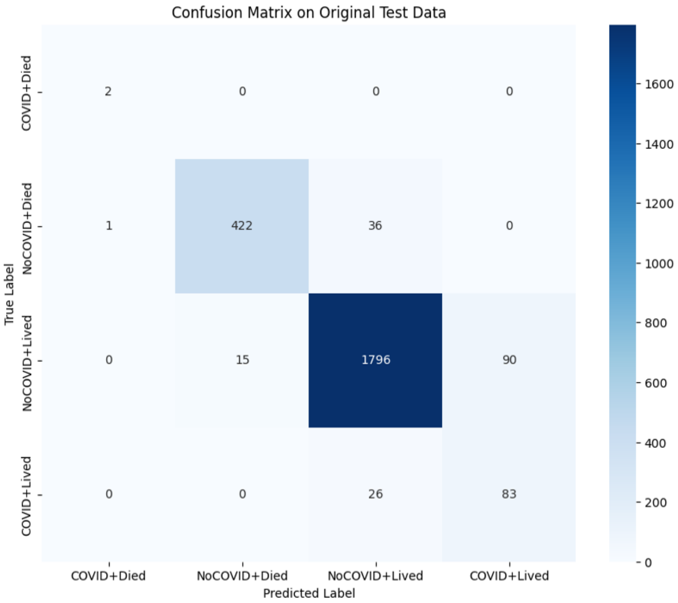
	COVID+Died	NoCOVID+Died	NoCOVID+Lived	COVID+Lived
COVID+Died	2	0	0	0
NoCOVID+Died	1	419	39	0
NoCOVID+Lived	0	39	1885	65
COVID+Lived	0	0	0	21

Fig. 18. Confusion Matrix of the ϵ = 1 Model with Original Data



	COVID+Died	NoCOVID+Died	NoCOVID+Lived	COVID+Lived
COVID+Died	1895	0	0	0
NoCOVID+Died	4	1805	90	11
NoCOVID+Lived	1	11	1808	88
COVID+Lived	0	3	59	1857

Fig. 19. Confusion Matrix of the ϵ = 3 Model with SMOTE data



	COVID+Died	NoCOVID+Died	NoCOVID+Lived	COVID+Lived
COVID+Died	2	0	0	0
NoCOVID+Died	1	422	36	0
NoCOVID+Lived	0	15	1796	90
COVID+Lived	0	0	26	83

Fig. 20. Confusion Matrix of the ϵ = 3 Model with Original Data