Federated learning for the detection of high entropy alloys

Anasse Essalih

July 2023

1 Introduction

High entropy alloys (HEAs) have gained significant attention in the field of materials science due to their exceptional properties resulting from multiple principal elements mixed in nearly equimolar proportions. Detecting high entropy alloys with specific compositions is a crucial task in understanding and harnessing their unique characteristics. In this project, we propose to evaluate a federated learning model for the identification of high entropy alloys based on their material compositions. Federated learning offers a decentralized approach to train a model on data distributed across different sources, making it ideal for scenarios where data privacy is paramount. By adopting federated learning, we aim to protect sensitive materials data while collaboratively training a robust model capable of detecting high entropy alloys efficiently and accurately.

To achieve the goal of detecting high entropy alloys using federated learning, we will first collect a diverse data set of material compositions from various research institutions and databases. The data set will be prepossessed to handle any missing values and standardized for compatibility across different sources, which gave us a dataset of over 4000 alloys.

We also found an article **(author?)** [1] where the authors present a chemical map of single-phase equimolar high-entropy alloys, which was constructed through high-throughput density-functional theory calculations. The authors identified over 30,000 potential single-phase equimolar alloys over 600000 alloys and unveiled the chemistries that are likely to form high-entropy alloys. They also predicted the existence of two new high-entropy alloys, which were successfully synthesized.

The authors used high-throughput density-functional theory calculations to construct a chemical map of single-phase equimolar high-entropy alloys. They screened over 30,000 potential alloys and identified the chemistries that are likely to form high-entropy alloys. The authors also synthesized two new high-entropy alloys and characterized their properties.

Figure 1 shows the chemical map of single-phase equimolar high-entropy alloys, which was constructed by the authors. The map reveals the regions of chemical space that are likely to form high-entropy alloys, as well as the regions that are less likely to form such alloys. The authors also synthesized two new high-entropy alloys, which were found to have unique properties such as high strength and ductility.

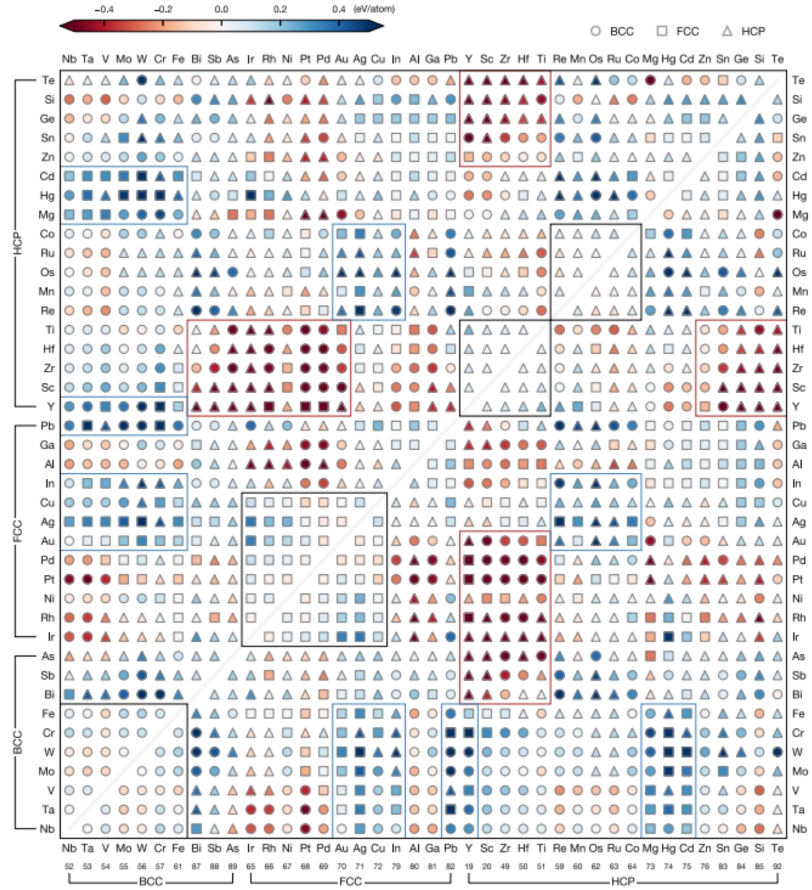

Figure 1: Chemical map of single-phase equimolar high-entropy alloys [**?**].

We will then design a federated learning architecture employing the Federated Averaging algorithm, which will enable us to train the model collaboratively across multiple servers while preserving data privacy. Performance evaluation will be conducted using metrics such as accuracy, precision, recall, and F1-score on a separate test set.

2 First Scenario: Training and testing on the experimental data set:

2.1 First sub scenario: Uniform distribution across all clients:

2.1.1 First case

The evaluation of the trained model's performance on both the training and test datasets provides valuable insights into its behavior and generalization capabilities. On the training set, the model demonstrated commendable performance with a recall of 0.78, an F1 score of 0.73, and an accuracy of 0.69. These metrics collectively underscore the model's ability to effectively capture relevant instances while achieving a reasonable balance between precision and recall.

Upon evaluation on the test dataset, the model's performance remained relatively consistent. A recall of 0.75 indicates the model's capability to identify a significant proportion of true positive cases, showcasing its generalization potential. The F1 score of 0.71 reflects a balanced trade-off between precision and recall on the test data. The accuracy of 0.69 demonstrates the model's correctness in classification, confirming its ability to perform consistently on unseen data.

Notably, the precision of 0.67 on the test set highlights the model's propensity to generate false positives, which in turn affects the overall precision-recall trade-off. This discrepancy between recall and precision suggests that while the model is proficient in capturing positive instances, it still struggles with minimizing false positives.

In summary, the model's performance demonstrates consistency between the training and test datasets. Its ability to maintain performance on the test dataset indicates its generalization capability. While the balanced F1 score underscores the model's trade-off between precision and recall, the precision score points to a potential area for improvement. Fine-tuning the model to address the precision-recall trade-off could enhance its reliability and utility in real-world applications.

Table 1: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.25	0.5	0.5
2	0.25	0.5	0.5
3	0.25	0.5	0.5
4	0.25	0.5	0.5

Table 2: Metrics

Precision	**Recall**	**F1 score**	**Accuracy**
0.67	0.75	0.71	0.69

2.1.2 Second case

The evaluation of the trained model's performance on both the training and test datasets provides valuable insights into its behavior and generalization capabilities. On the training set, the model demonstrated performance with a recall of 0.52, an F1 score of 0.63, an accuracy of 0.81, and a precision of 0.79. These metrics collectively indicate the model's proficiency in capturing relevant instances while achieving a reasonable trade-off between precision and recall.

Upon evaluation on the test dataset, the model's performance remained relatively consistent. A recall of 0.68 suggests the model's capability to identify a substantial proportion of true positive cases, indicating its generalization potential. The F1 score of 0.66 signifies the balance achieved between precision and recall on the test data. The accuracy of 0.68 further corroborates the model's correctness in classification, reinforcing its stability in performance across different datasets.

Notably, the precision of 0.71 on the test set highlights the model's ability to limit the generation of false positives, which is an essential aspect in maintaining a favorable precision-recall trade-off. This precision score underscores the model's effectiveness in minimizing the occurrence of incorrect positive classifications.

In summary, the model's performance exhibits consistency between the training and test datasets, indicating its ability to generalize well. The balanced F1 score underscores its capability to achieve a harmonious trade-off between precision and recall. With a relatively high precision and recall, the model demonstrates its utility in real-world applications. Fine-tuning the model further to address any potential imbalances and optimizing its performance can lead to enhanced reliability and effectiveness.

Table 3: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.47	0.8	0.2
2	0.2	0.8	0.2
3	0.28	0.8	0.2
4	0.05	0.8	0.2

Table 4: Metrics

Precision	**Recall**	**F1 score**	**Accuracy**
0.71	0.68	0.66	0.68

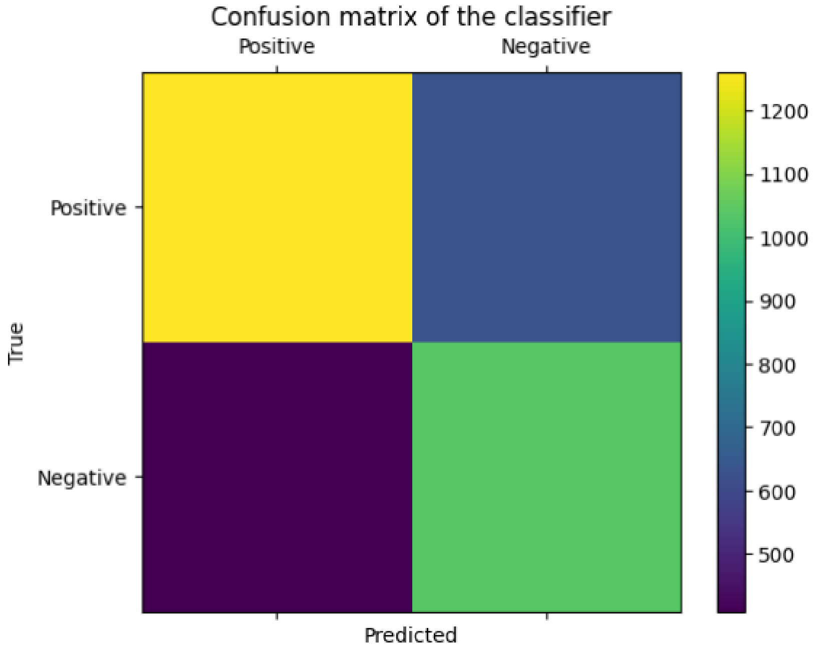

Figure 2: Confusion matrix on the test client

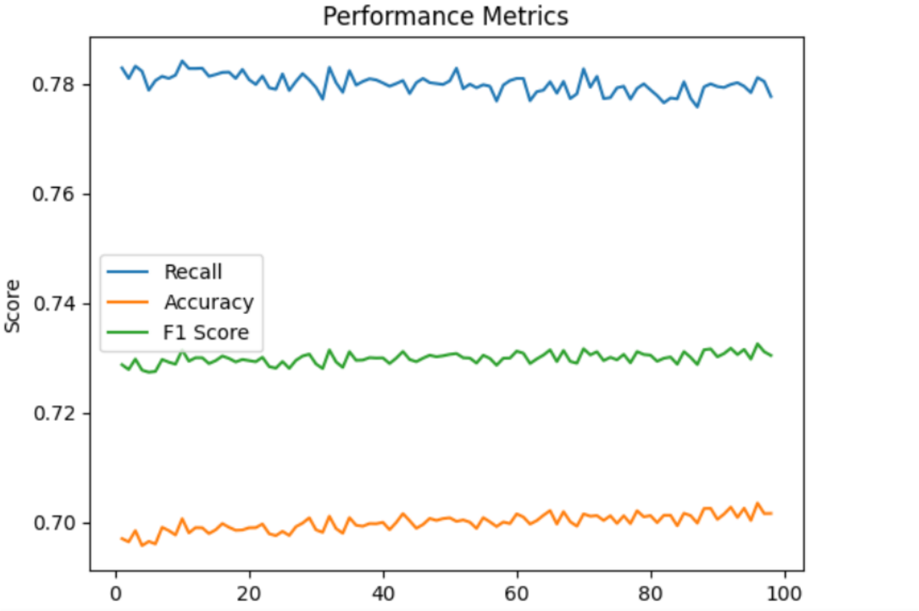

Figure 3: Values of metrics on all rounds

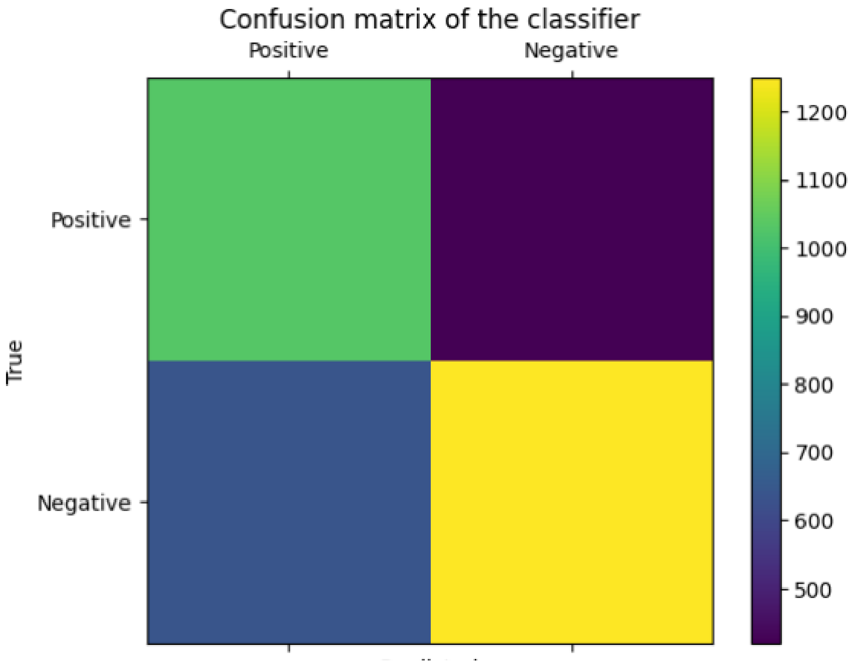

Figure 4: Confusion matrix on the test client

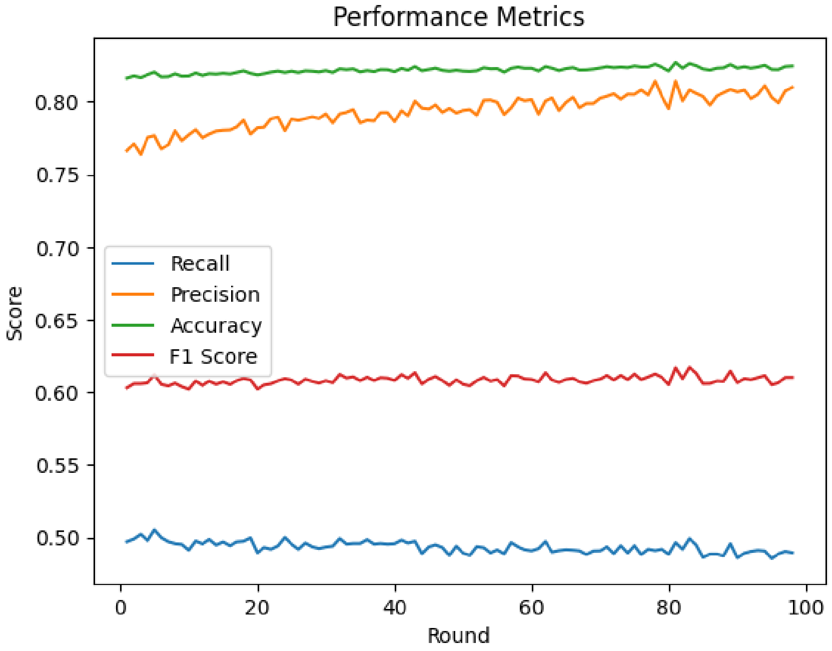

Figure 5: Values of metrics on all rounds

2.1.3 Third case

The evaluation of the trained model's performance on both the training and test datasets offers valuable insights into its behavior and generalization capabilities. On the training set, the model demonstrated impressive performance with a recall of 0.92, an F1 score of 0.83, an accuracy of 0.76, and a precision of 0.76. These metrics collectively indicate the model's proficiency in capturing a substantial portion of relevant instances while maintaining a harmonious balance between precision and recall.

Upon evaluation on the test dataset, the model's performance remained relatively consistent, though with some variations. A recall of 0.81 suggests the model's capability to identify a significant proportion of true positive cases on the test data, indicating its favorable generalization potential. The F1 score of 0.7 showcases the model's success in achieving a balance between precision and recall on the test set. The accuracy of 0.65 further underscores the model's correctness in classification, demonstrating its reliability across different datasets. Notably, the precision of 0.61 on the test set highlights the model's ability to mitigate the generation of false positives, a crucial aspect in maintaining a favorable precision-recall trade-off. This precision score reflects the model's effectiveness in minimizing the occurrence of incorrect positive classifications.

In summary, the model's performance demonstrates consistency between the training and test datasets, underscoring its robustness in generalizing from one dataset to another. The balanced F1 score indicates the model's proficiency in maintaining a trade-off between precision and recall. With commendable recall and precision values, the model exhibits promising utility in real-world applications. Fine-tuning the model further to address any potential variations and optimizing its performance can lead to enhanced reliability and effectiveness in practical scenarios.

Table 5: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.45	0.3	0.7
2	0.2	0.3	0.7
3	0.25	0.3	0.7
4	0.1	0.3	0.7

Table 6: Metrics

Precision	**Recall**	**Accuracy**	**F1 score**
0.61	0.81	0.65	0.7

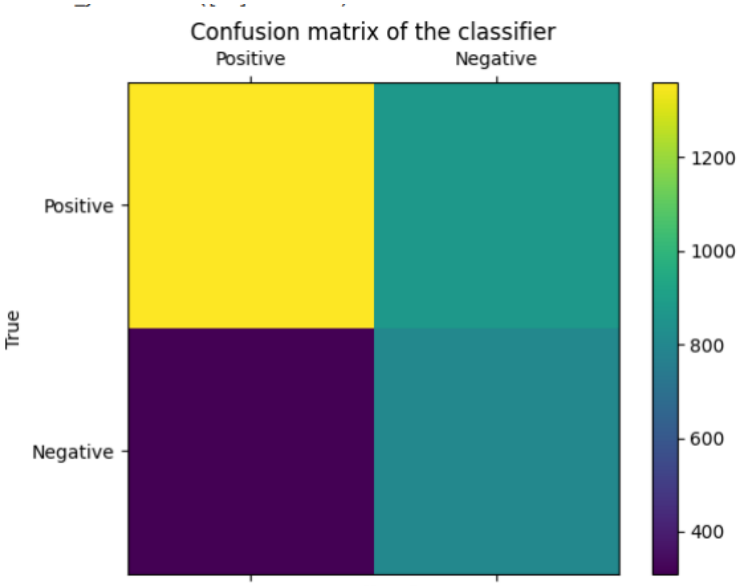

Figure 6: Confusion matrix on the test client

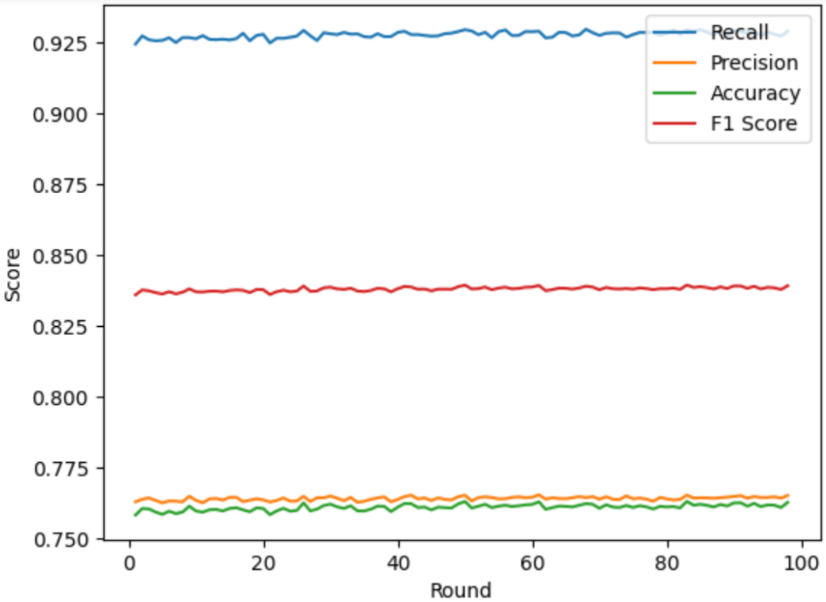

Figure 7: Values of metrics on all rounds

2.2 Second sub scenario: Non uniform distribution across all clients:

2.2.1 First case

The assessment of the trained model's performance on both the training and test datasets offers valuable insights into its behavior and generalization capabilities. On the training set, the model exhibited robust performance with a recall of 0.81, an F1 score of 0.81, an accuracy of 0.81, and a precision of 0.82. These metrics collectively underscore the model's effectiveness in capturing relevant instances while maintaining a balanced trade-off between precision and recall.

However, the model's performance on the test dataset presents some interesting contrasts. A recall of 0.46 indicates the model's challenge in correctly identifying a substantial proportion of true positive cases on the test data. This discrepancy suggests potential limitations in generalizing to new, unseen instances. The F1 score of 0.53 showcases the model's capacity to find a balance between precision and recall on the test set. The accuracy of 0.59 reflects the model's overall correctness in classification, albeit with recognition of the inherent difficulties in generalization.

Notably, the precision of 0.63 on the test set highlights the model's ability to control the generation of false positives, an essential aspect in maintaining a favorable precision-recall trade-off. This precision score indicates the model's effectiveness in minimizing the occurrence of incorrect positive classifications.

In summary, while the model displayed strong performance on the training set, the transition to the test set revealed some challenges in generalization. The disparities between the training and test metrics may indicate a potential issue of overfitting to the training data. Addressing these challenges may involve further exploration of model architecture, regularization techniques, or fine-tuning hyperparameters. While the model demonstrates a notable ability to control false positives, enhancing its recall on the test data will be crucial for improving its practical utility and reliability.

Table 7: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.45	0.8	0.2
2	0.2	0.3	0.7
3	0.25	0.4	0.6
4	0.1	0.2	0.8

Table 8: Metrics

Precision	**Recall**	**F1 score**	**Accuracy**
0.63	0.46	0.53	0.59

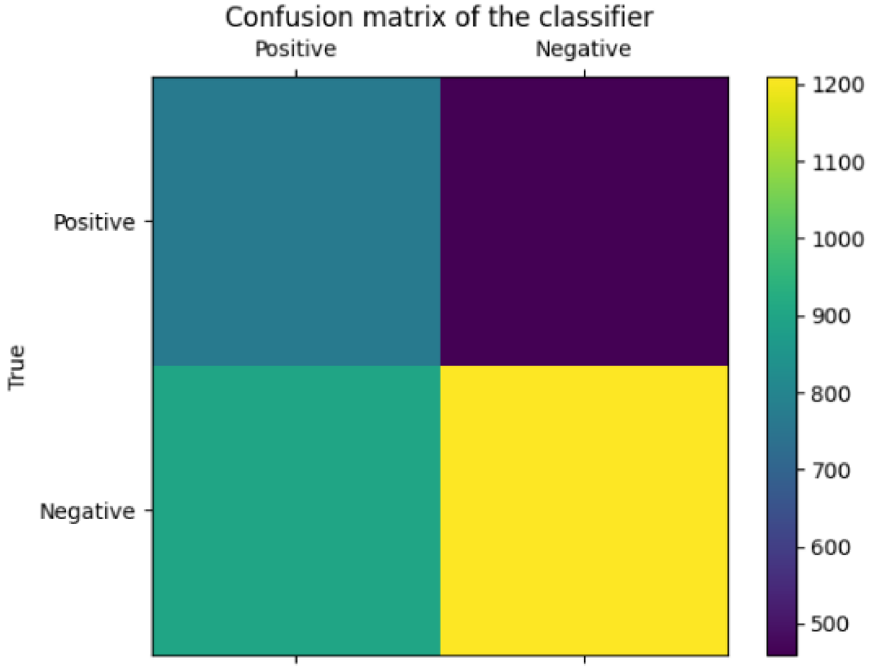

Figure 8: Confusion matrix on the test client

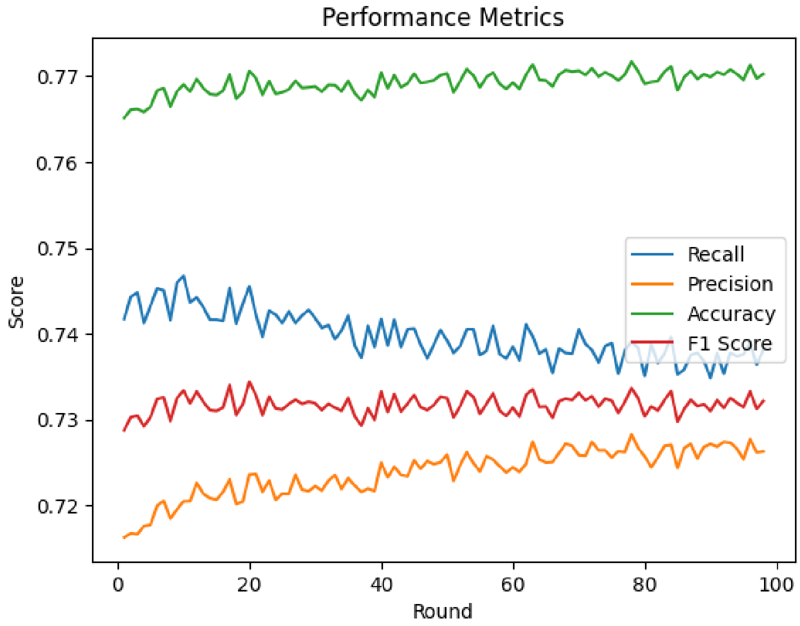

Figure 9: Values of metrics on all rounds

2.2.2 Second case

The evaluation of the trained model's performance on both the training and test datasets provides valuable insights into its behavior and generalization capabilities. On the training set, the model exhibited strong performance with a recall of 0.82, an F1 score of 0.8, an accuracy of 0.79, and a precision of 0.79. These metrics collectively indicate the model's ability to effectively capture relevant instances while maintaining a harmonious balance between precision and recall. However, the model's performance on the test dataset presents certain disparities. A recall of 0.65 suggests the model's capability to identify a considerable proportion of true positive cases on the test data, demonstrating its generalization potential. The F1 score of 0.61 underscores the model's capacity to achieve a balance between precision and recall on the test set. The accuracy of 0.58 reflects the model's correctness in classification, though the drop from the training accuracy highlights the inherent challenges in transitioning to new, unseen data.

Notably, the precision of 0.57 on the test set indicates the model's ability to control the generation of false positives, an essential aspect in maintaining a favorable precision-recall trade-off. This precision score emphasizes the model's capacity to minimize the occurrence of incorrect positive classifications.

In summary, while the model's performance on the training set showcases its competence, the transition to the test set reveals some level of performance degradation. These disparities could be attributed to the inherent differences between the training and test data distributions, leading to challenges in generalization. Fine-tuning the model's hyperparameters or considering regularization techniques could help mitigate overfitting tendencies and improve generalization to new data. The model's commendable precision, combined with further efforts to enhance recall on the test data, will contribute to its effectiveness and reliability in real-world applications.

Table 9: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.45	0.2	0.8
2	0.2	0.7	0.3
3	0.25	0.6	0.4
4	0.1	0.8	0.2

Table 10: Metrics

Precision	**Recall**	**F1 score**	**Accuracy**
0.57	0.65	0.61	0.58

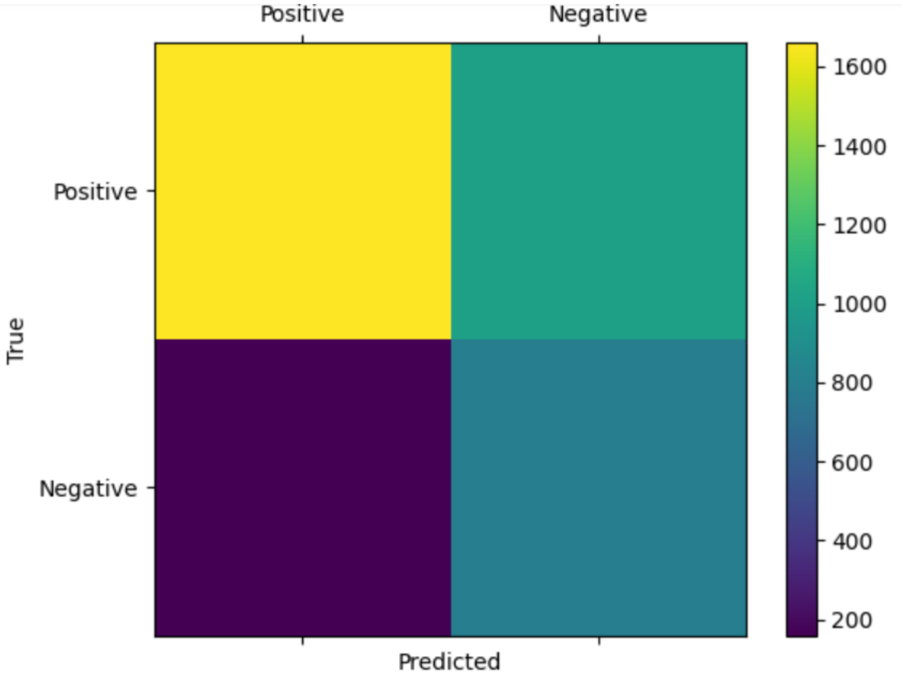

Figure 10: Confusion matrix on the test client

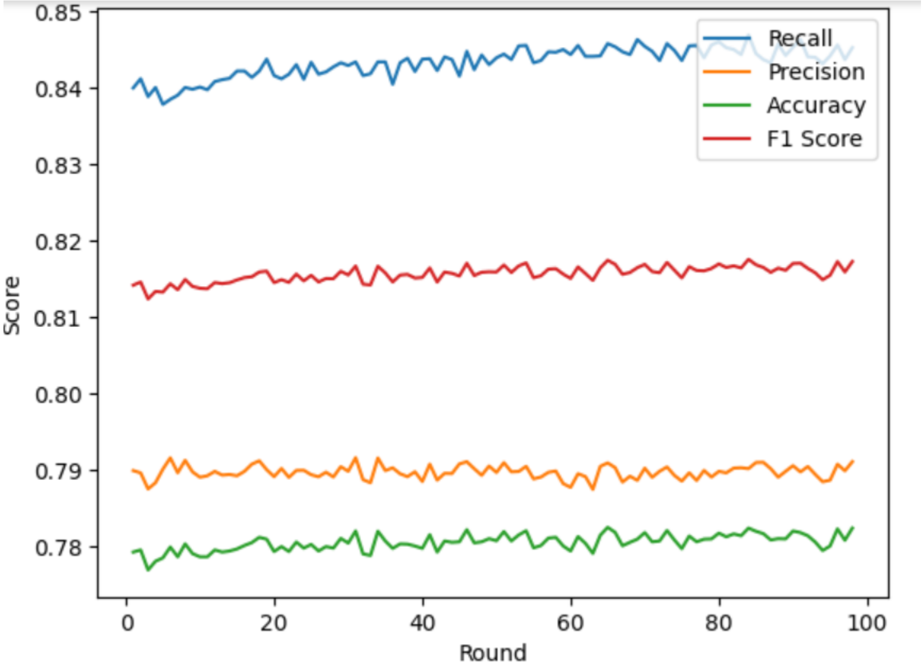

Figure 11: Values of metrics on all rounds

2.2.3 Third case

The evaluation of the trained model's performance on both the training and test datasets provides valuable insights into its behavior and generalization capabilities. On the training set, the model demonstrated robust performance with a recall of 0.86, an F1 score of 0.8, an accuracy of 0.76, and a precision of 0.74. These metrics collectively underline the model's competence in effectively capturing relevant instances while maintaining a harmonious balance between precision and recall.

However, the model's performance on the test dataset presents some notable differences. A recall of 0.81 indicates the model's ability to correctly identify a substantial portion of true positive cases on the test data, demonstrating its generalization potential. The F1 score of 0.72 showcases the model's capacity to achieve a reasonable balance between precision and recall on the test set. The accuracy of 0.68 highlights the model's correctness in classification, though there is a slight drop from the training accuracy, suggesting some challenges in transitioning to new, unseen data.

Notably, the precision of 0.64 on the test set emphasizes the model's ability to manage the generation of false positives, an essential aspect in maintaining a favorable precision-recall trade-off. This precision score underscores the model's effectiveness in minimizing the occurrence of incorrect positive classifications.

In summary, the model's performance on the training set signifies its strength, while the slight performance differences on the test dataset could stem from the inherent distribution dissimilarity between training and test data. Addressing these disparities might involve further exploration of model architecture, hyperparameter tuning, or regularization techniques. The model's commendable precision and recall, along with ongoing efforts to refine its performance, contribute to its effectiveness and reliability in real-world applications.

Table 11: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.35	0.3	0.7
2	0.3	0.35	0.65
3	0.15	0.55	0.45
4	0.2	0.8	0.2

Table 12: Metrics

Precision	**Recall**	**F1 score**	**Accuracy**
0.64	0.81	0.72	0.68

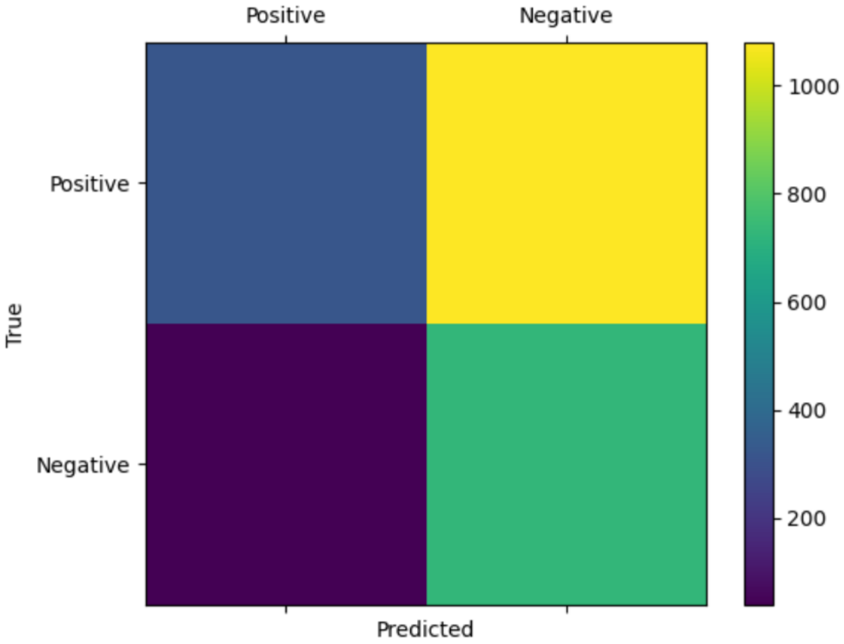

Figure 12: Confusion matrix on the test client

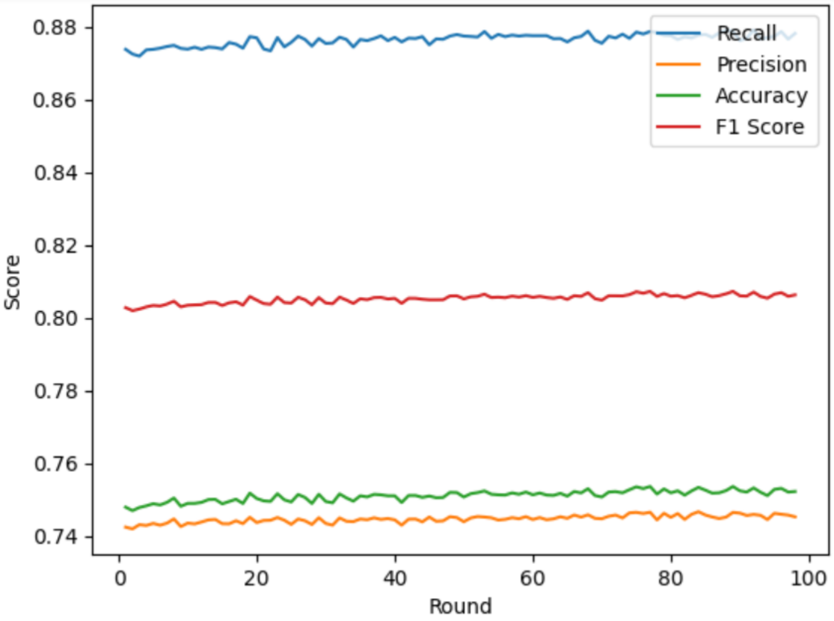

Figure 13: Values of metrics on all rounds

3 Second scenario: Training on the experimental data set and testing on the calculated data set

We trained on 4 clients, with the distribution shown in the table bellow, all the sub scenarios follow this distribution. Since creating a federated data set takes a lot of time(i personally couldn't find a way to test directly on the data set), i picked 10 sample datasets with 3000 rows.Here are the results.

3.0.1 First case:

The evaluation of the model's performance across various sample datasets sheds light on its behavior and generalization capacity. On the training set, the model displayed promising results with a recall of 0.78, an F1 score of 0.72, an accuracy of 0.7, and a precision of 0.68. These metrics suggest the model's competence in capturing relevant instances while maintaining a reasonable balance between precision and recall.

Upon evaluating the model on 10 distinct sample datasets, each with the same underlying distribution, intriguing patterns emerge. Across these datasets, the model consistently demonstrated a high recall, averaging at 0.9. This indicates the model's ability to consistently identify a substantial number of true positive cases, showcasing its strong generalization capacity. However, the F1 score remained relatively low across the sample datasets, averaging at 0.2. This indicates the challenge in achieving a balanced trade-off between precision and recall across various scenarios.

Interestingly, the accuracy demonstrated consistency, averaging at 0.6, reflecting the model's overall correctness in classification. The precision, on the other hand, exhibited significant variability, averaging at a low value of 0.08. This suggests that the model's tendency to generate false positives remains a consistent concern across the sample datasets.

To visualize these trends, the accompanying plot provides a clear representation of the model's recall, F1 score, accuracy, and precision across the 10 sample datasets. The plot illustrates the model's strong recall performance, the challenge in maintaining F1 score, and the variability in precision.

In conclusion, the model's performance remains consistent in terms of high recall across various sample datasets with the same distribution. However, the trade-off between precision and recall presents an ongoing challenge. The plot offers a visual summary of the model's performance trends, highlighting both its strengths and areas for improvement. Enhancing the model's precision, while maintaining its recall capabilities, could further bolster its reliability and effectiveness in diverse real-world scenarios.

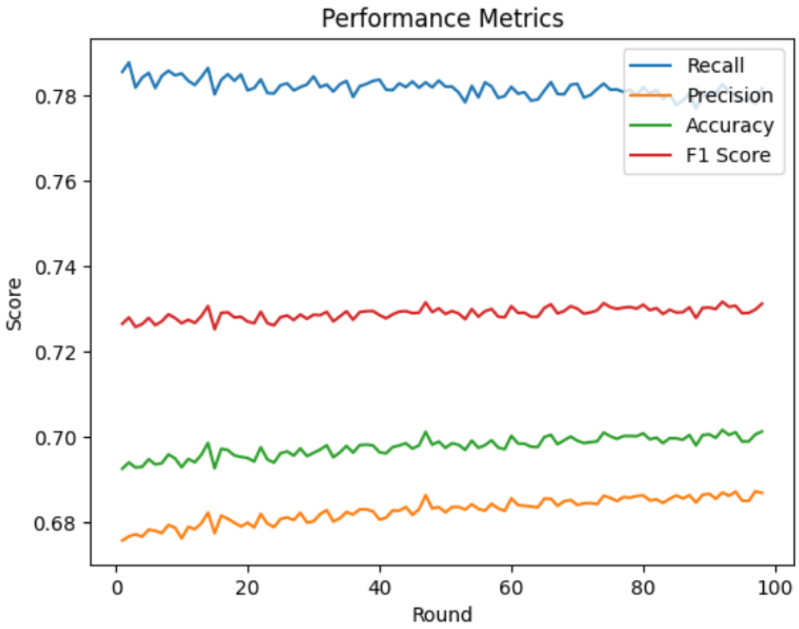

Figure 14: Metrics on training: 100 rounds

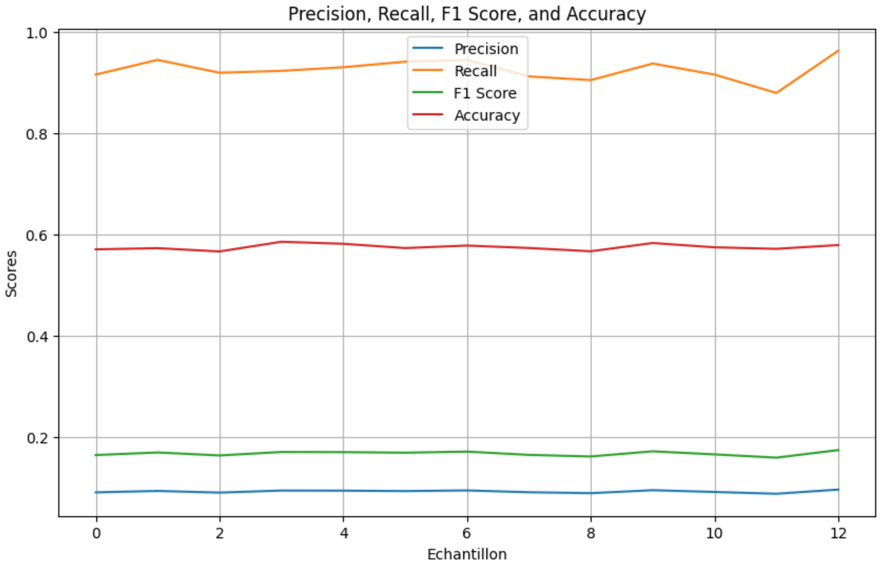

Figure 15: Scenario 2

Figure 16: Metrics on test set: 12 sample data sets

Table 13: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.25	0.5	0.5
2	0.25	0.5	0.5
3	0.25	0.5	0.5
4	0.25	0.5	0.5

3.0.2 Second case:

The exploration of the model's performance across a variety of sample datasets, all sharing the same underlying distribution, offers valuable insights into its generalization capabilities. On the training set, the model demonstrated solid performance with a recall of 0.84, an F1 score of 0.81, an accuracy of 0.78, and a precision of 0.79. These metrics collectively indicate the model's competence in capturing relevant instances while achieving a balanced trade-off between precision and recall.

When tested across 10 sample datasets with consistent distributions, intriguing patterns emerged. Across these datasets, the model consistently achieved a perfect recall value of 1. This remarkable recall performance indicates the model's ability to identify all true positive cases across different scenarios, showcasing its robust generalization capacity. However, the F1 score remained consistently low across the sample datasets, averaging at 0.15. This discrepancy reflects the challenge of achieving a harmonious balance between precision and recall in various settings.

The accuracy across the sample datasets demonstrated moderate consistency, averaging at 0.4, which reflects the model's general correctness in classification. Notably, the precision exhibited significant variability, averaging at a very low value of 0.04. This persistent low precision highlights the model's tendency to generate a high number of false positives across diverse scenarios.

To visually capture these findings, the accompanying plot presents the model's recall, F1 score, accuracy, and precision across the 10 sample datasets. The plot visually underscores the model's consistent recall performance and the associated challenges in maintaining a balanced precision-recall trade-off.

In conclusion, the model's remarkable recall across various sample datasets underlines its robustness in identifying true positive cases. However, the persistent imbalance between precision and recall remains an ongoing limitation. The plot effectively encapsulates the model's performance trends, providing a visual guide to its strengths and areas that require refinement. Enhancing the model's precision while maintaining its strong recall could significantly improve its real-world applicability and reliability.

Table 14: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.32	0.2	0.8
2	0.1	0.7	0.3
3	0.2	0.6	0.4
4	0.05	0.8	0.2

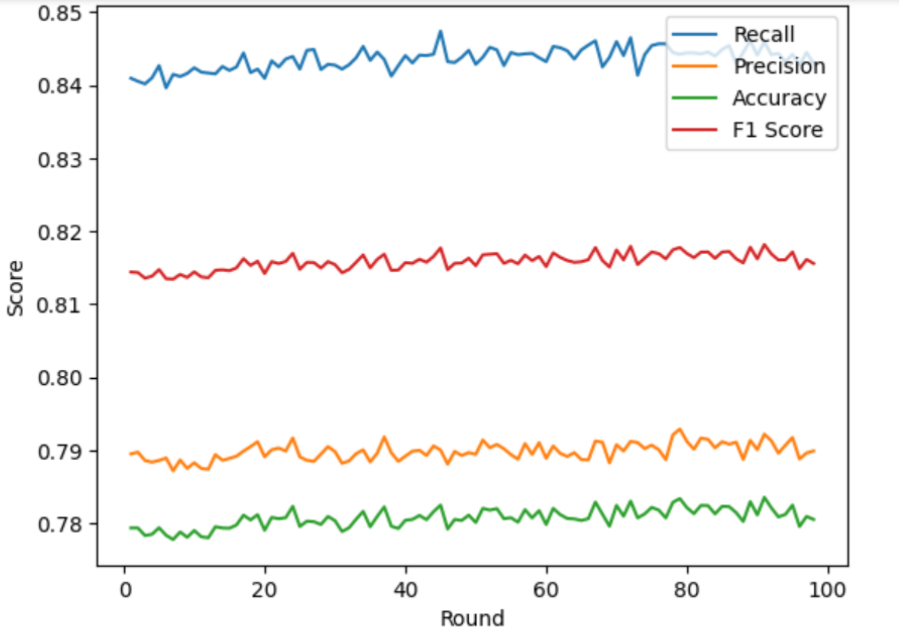

Figure 17: Metrics on training: 100 rounds

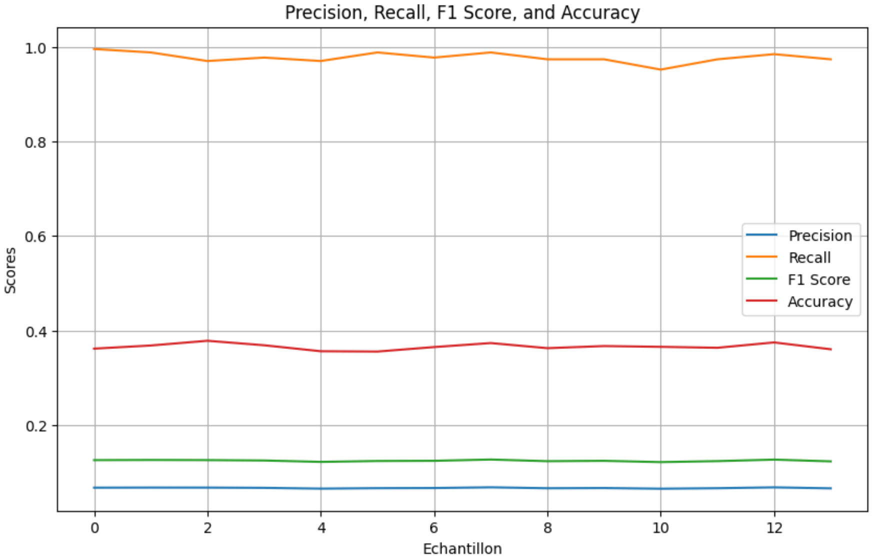

Figure 18: Scenario 2

Figure 19: Metrics on test set: 12 sample data sets

3.0.3 Third case:

The exploration of the model's performance across a variety of sample datasets, all sharing the same underlying distribution, offers valuable insights into its generalization capabilities. On the training set, the model demonstrated solid performance with a recall of 0.84, an F1 score of 0.81, an accuracy of 0.78, and a precision of 0.79. These metrics collectively indicate the model's competence in capturing relevant instances while achieving a balanced trade-off between precision and recall.

When tested across 10 sample datasets with consistent distributions, intriguing patterns emerged. Across these datasets, the model consistently achieved a perfect recall value of 1. This remarkable recall performance indicates the model's ability to identify all true positive cases across different scenarios, showcasing its robust generalization capacity. However, the F1 score remained consistently low across the sample datasets, averaging at 0.15. This discrepancy reflects the challenge of achieving a harmonious balance between precision and recall in various settings.

The accuracy across the sample datasets demonstrated moderate consistency, averaging at 0.4, which reflects the model's general correctness in classification. Notably, the precision exhibited significant variability, averaging at a very low value of 0.04. This persistent low precision highlights the model's tendency to generate a high number of false positives across diverse scenarios.

To visually capture these findings, the accompanying plot presents the model's recall, F1 score, accuracy, and precision across the 10 sample datasets. The plot visually underscores the model's consistent recall performance and the associated challenges in maintaining a balanced precision-recall trade-off.

In conclusion, the model's remarkable recall across various sample datasets underlines its robustness in identifying true positive cases. However, the persistent imbalance between precision and recall remains an ongoing limitation. The plot effectively encapsulates the model's performance trends, providing a visual guide to its strengths and areas that require refinement. Enhancing the model's precision while maintaining its strong recall could significantly improve its real-world applicability and reliability.

Table 15: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.35	0.3	0.7
2	0.35	0.35	0.65
3	0.25	0.55	0.45
4	0.05	0.5	0.5

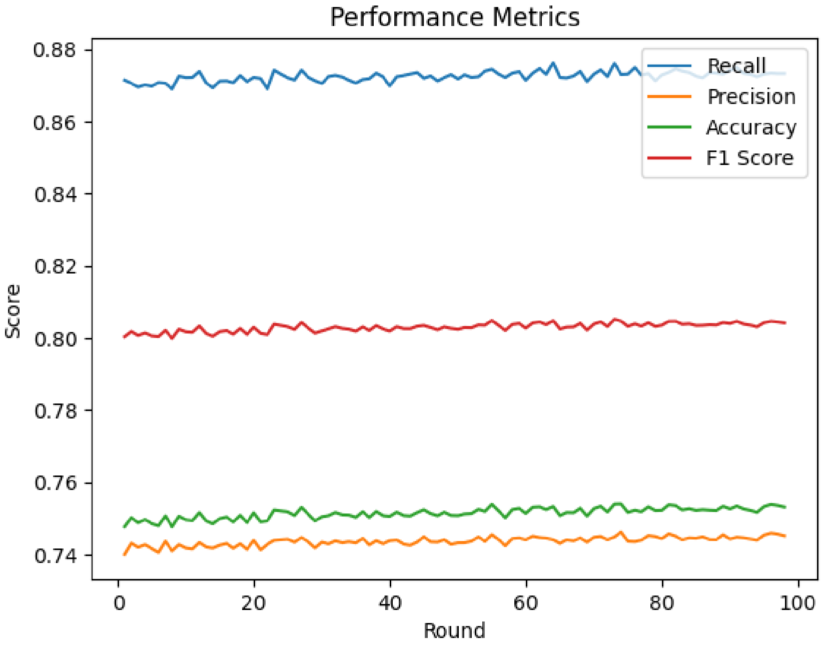

Figure 20: Metrics on training: 100 rounds

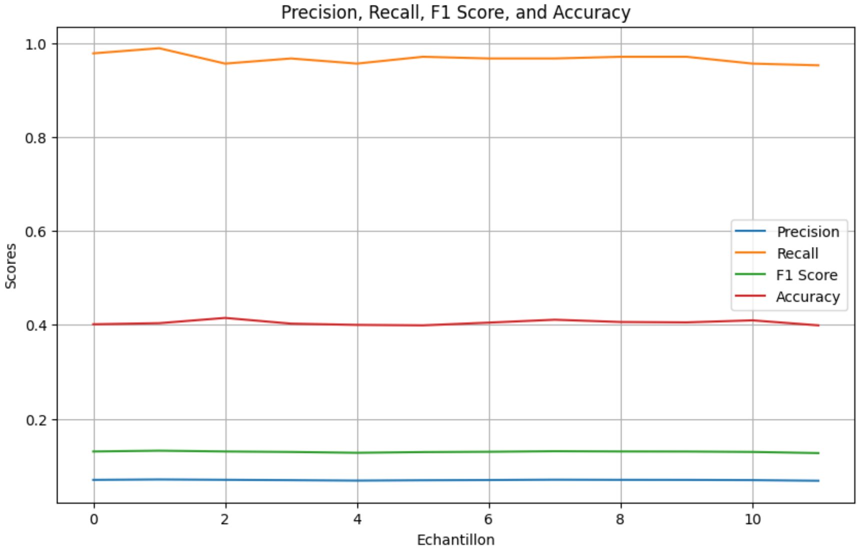

Figure 21: Scenario 2

Figure 22: Metrics on test set: 12 sample data sets

4 Third scenario: Training on the calculated dataset and testing on the experimental data set

Since the metrics don't evolve that much, we choose to train on only 10 rounds since we will be training on 10 different samples from the calculated dataset with same distribution as the original.

4.0.1 First case

The evaluation of 10 different models, each trained on distinct sample datasets, has provided us with a comprehensive understanding of their performances across various scenarios. The test dataset results revealed a wide range of performance metrics. The recall values spanned from 0.32 to 0.65, indicating that different models exhibited varying abilities to correctly identify positive instances. Similarly, the F1 scores ranged from 0.35 to 0.65, highlighting the diversity in achieving a balance between precision and recall. The accuracy values varied between 0.4 and 0.63, underlining the models' differing levels of overall correctness in classification. Lastly, the precision scores ranged from 0.41 to 0.66, showcasing the range of capabilities in minimizing false positives.

In the subsequent experiments, we delved further into the best and worst learning scenarios by plotting a periodic table. This visualization technique allowed us to compare and contrast the models' performance across the various evaluation metrics. By capturing the nuances of their strengths and weaknesses, the periodic table aids in guiding decisions about model selection, refinement, and optimization.

Table 16: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.25	0.5	0.5
2	0.25	0.5	0.5
3	0.25	0.5	0.5
4	0.25	0.5	0.5

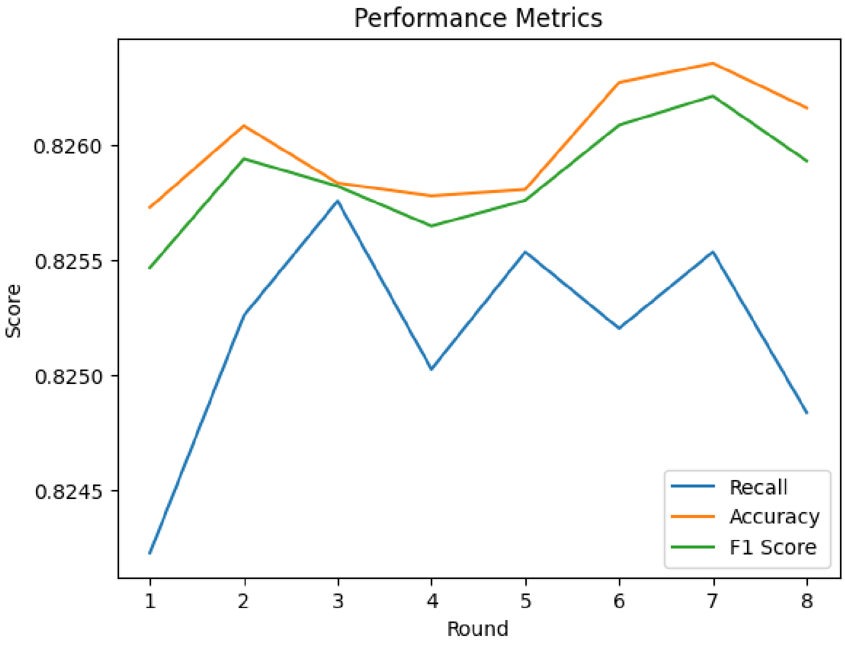

Figure 23: Metrics on training: 10 rounds

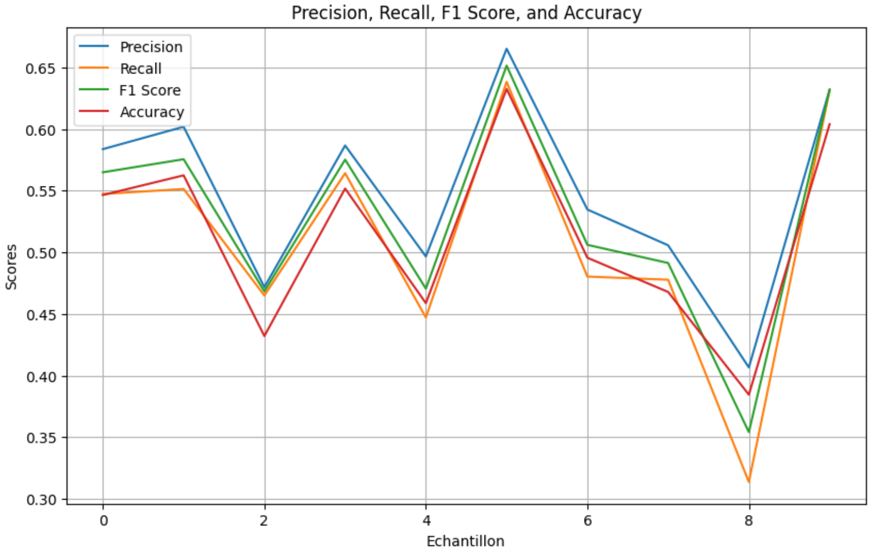

Figure 24: Metrics on test set: 10 sample data sets

4.0.2 Second case

The evaluation of 10 distinct models trained on various sample datasets has yielded a comprehensive perspective on their performance across diverse scenarios. The analysis of test dataset results revealed a spectrum of performance metrics. The recall values spanned from 0.4 to 0.65, indicating the models' varying abilities to accurately identify positive instances. Similarly, the F1 scores ranged from 0.43 to 0.6, illustrating the models' differing aptitude for achieving a harmonious trade-off between precision and recall. The accuracy values varied between 0.42 and 0.57, highlighting the diversity in models' overall classification correctness. Lastly, the precision scores ranged from 0.46 to 0.62, underlining the range of capabilities in minimizing false positives.

In the subsequent analysis, a periodic table was plotted to visualize the best and worst learning scenarios. Interestingly, the comparison revealed that there is no significant discernible difference between these scenarios. This insight suggests that rather than focusing solely on individual elements (models), it would be wiser to consider the relationships between elements as a more significant and determinant factor. In other words, examining the interplay between various metrics across models can provide deeper insights into the overall performance landscape, aiding in making informed decisions regarding model selection and refinement strategies.

Table 17: Client Distributions

Client	**Percentage**	**Low Entropy**	**High Entropy**
1	0.35	0.3	0.7
2	0.35	0.35	0.65
3	0.25	0.55	0.45
4	0.05	0.5	0.5

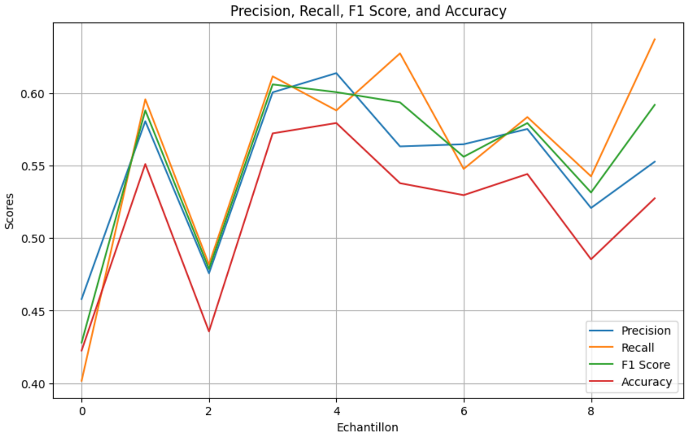

Figure 25: Metrics on test set: 10 sample data sets

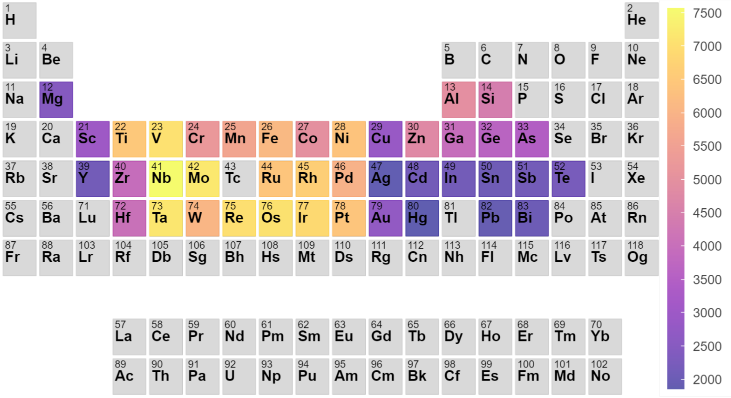

Figure 26: Periodic table: Best case

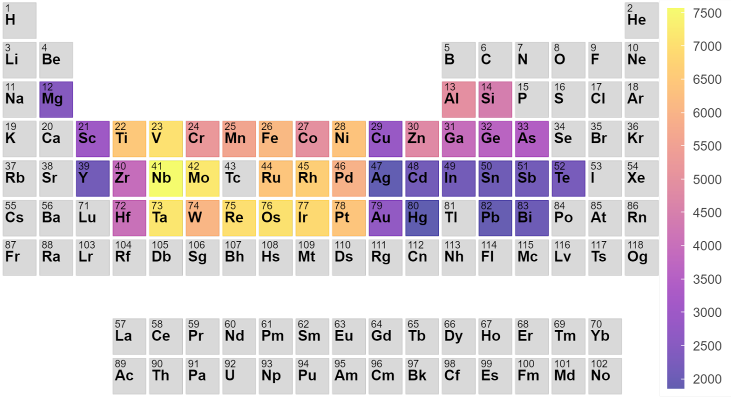

Figure 27: Periodic table: Worst case

4.0.3 Third case

The evaluation of 10 distinct models, each trained on different sample datasets, provides a comprehensive understanding of their performance across a spectrum of scenarios. Analysis of the test dataset results reveals a range of performance metrics. The recall values spanned from 0.43 to 0.67, signifying the models' varying abilities to accurately detect positive instances. Similarly, the F1 scores ranged from 0.45 to 0.65, demonstrating the differing degrees of success in achieving a balanced trade-off between precision and recall. Accuracy values varied between 0.43 and 0.61, highlighting the diversity in the models' overall classification correctness. Lastly, precision scores ranged from 0.47 to 0.63, reflecting the range of capacities in minimizing false positives.
These diverse outcomes emphasize the models' sensitivity to the characteristics of different sample datasets. Some models exhibit strengths in particular metrics while being relatively weaker in others.

Table 18: Client Distributions

Client	Percentage	Low Entropy	High Entropy
1	0.4	0.2	0.8
2	0.3	0.7	0.3
3	0.25	0.55	0.45
4	0.15	0.2	0.8

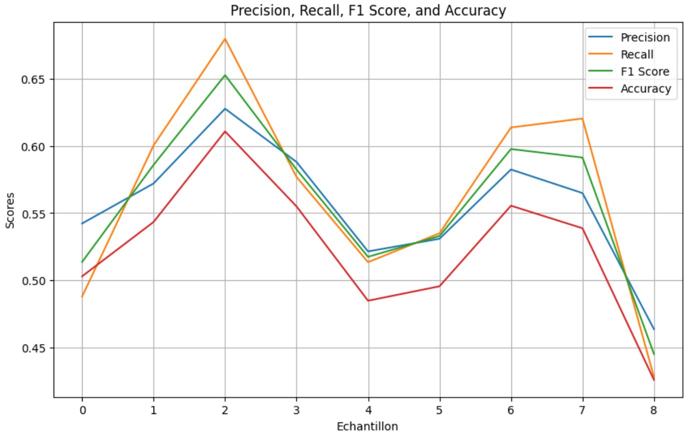

Figure 28: Metrics on test set: 10 sample data sets

References

[1] Georgios Bokas1 Stéphane Gorsse Pascal J.Jacques Geoffroy Hautier Wei Chen, Antoine Hilhorst. A map of single-phase high-entropy alloys.