International Journal of Electrical and Electronics Engineers
Volume 18, Issue No. 01, Jan-June 2026
ISSN (0) 2321-2055 ISSN (P) 2321-2045

AN UNCERTAINTY AWARE DEEP LEARNING FRAMEWORK FOR ACTIVE LEARNING WITH MISSING DATA

Maheswari Kannedari
department of ECE
Tirumala Engineering College
maheswarikannedari07@gmail.com

Kavya Ambati
department of ECE
Tirumala Engineering College
kavyaambati1106@gmail.com

Anushka Reddy Beeram
department of ECE
Tirumala Engineering College
anushkabeeram04@gmail.com

Neelima Anumala
department of ECE
Tirumala Engineering College
anumalaneelimal@gmail.com

Venkata Hanuman
department of ECE
Tirumala Engineering College
venkatahanuman@gmail.com

***Abstract*— In many real-world situations, datasets have missing or incomplete values because of sensor failures, mistakes in data collection, or privacy concerns. If not handled correctly, missing data can make predictive models much less accurate and reliable. People often use traditional methods like mean, median, or mode imputation to solve this problem, but these methods don't always show how features are related to each other in a complex way and can cause information loss. So, you need more advanced methods to learn from datasets that aren't complete. This project suggests a deep learning framework that takes into account uncertainty for active learning with missing data. The framework enhances model performance in the presence of incomplete data by integrating deep learning models with uncertainty estimation and active learning strategies. A neural network-based architecture is utilized to extract significant patterns from datasets, even in the presence of missing input values. The model assesses prediction uncertainty to pinpoint data samples where confidence is lacking. Active learning is employed to minimize the necessity for extensive labeled datasets by selecting only the most informative samples for labeling. By concentrating on uncertain or informative data points, the system enhances training efficiency while decreasing labeling costs. The proposed framework is tested using structured datasets with missing values, such as the Titanic dataset. Experimental results indicate that the integration of deep learning, uncertainty estimation, and active learning enhances prediction accuracy and model reliability compared to conventional methods for handling missing data. This method is a good way to build strong machine learning systems. The suggested method is a good way to build strong machine learning systems that can work with incomplete data in real-world situations.**

I. INTRODUCTION

Data annotation has become a time-consuming and expensive process because unlabeled data is growing so quickly, especially from IoT devices. It's not always possible to label large datasets, which makes it hard to build good machine learning models. Active Learning (AL) solves this problem by only choosing the most useful data points to label. This cuts costs while still letting models generalize well.

In Active Learning, a model asks for specific samples from an unlabeled dataset and adds them to the training set. Adding more queries makes performance better, but it also makes labeling more expensive. So, the goal is to get high performance with a small query budget by focusing on the samples that give the most information.

Most AL methods, on the other hand, assume that datasets are complete. This isn't true in the real world, where missing values are common because of mistakes made when collecting or storing data. Before using AL, these missing values need to be dealt with, usually by using imputation methods. Existing imputation methods, particularly those based on deep learning, necessitate extensive datasets and may yield inaccurate outcomes when labeled data is scarce. This adds uncertainty to the imputation process, which can make the model work worse by adding noise or misleading data points.

Some earlier studies looked at imputation uncertainty, but they had some problems, like choosing the first labeled data randomly, using only one imputation method, and not having enough evaluation metrics. The proposed Active Learning with Missing Data (ALMD) approach presents an enhanced framework to address these challenges. The suggested method has two parts: exploration and exploitation. The exploration phase finds samples that are typical of the whole data space, while the exploitation phase focuses on samples that are near decision boundaries and give useful information. This makes the performance better, especially in datasets that aren't balanced.

The model also uses a new multiple imputation method that picks the most important features for estimating missing values. This lowers the prediction error and makes the model more accurate. It also uses more than one evaluation metric to give a more complete picture of performance.

Experimental outcomes across diverse datasets, encompassing balanced, imbalanced, binary, and multiclass scenarios, illustrate the efficacy of the proposed approach relative to current active learning methodologies.

II. LITERATURE REVIEW

A. Tharwat and W. Schenck (2020) This paper suggests an active learning method for datasets that aren't balanced by balancing exploration and exploitation strategies. By choosing samples that give useful information, it makes classification more accurate. But it doesn't deal with missing data or uncertainty very well.

W. Liu et al. (2021)The authors created a framework for active learning that works with multi-class imbalanced data

streams that change over time. The model changes as the data distributions change over time. But it mostly talks about data imbalance and doesn't address missing data problems.

A. Tharwat (2021)This study shows different ways to measure how well a classification works, such as accuracy, precision, recall, and F1-score. It helps you figure out how well the model is working. But it doesn't help with how to deal with missing data.

C. K. Enders (2022)This work elucidates various statistical techniques for managing missing data, including imputation methods. It gives a strong base in theory. But it doesn't work with modern deep learning methods.

A. Tharwat and W. Schenck (2022)The authors put forward a low-query active learning technique that uses pseudo-labeling to cut down on the cost of labeling. It makes things more efficient by picking out important samples. But it doesn't take into account missing data or figuring out how sure you are.

Z. Chai et al. (2022)This paper presents a profound probabilistic transfer learning framework for managing absent data. Probabilistic modeling makes predictions more accurate. The method, on the other hand, is hard to compute and uses a lot of resources.

M. P. Śmieja et al. (2022)The authors put forward MisConv, a convolutional neural network made to work with missing data directly. It makes learning better without needing to be preprocessed. But it doesn't include figuring out how likely something is to happen.

A. Tharwat and W. Schenck (2024/2025)This research concentrates on the integration of active learning with the management of missing data. It makes the model work better by picking samples that are useful. But it doesn't have a very good way of dealing with uncertainty.

A. Tharwat et al. (Various Contributions)The authors contributed to active learning, optimization, and classification methods in several studies. Their work makes the model work better and faster. But there are still limits to how well deep learning, uncertainty, and missing data can all work together.

III. DATASET DESCRIPTION

The dataset used in this study is the Titanic dataset, a widely recognized benchmark in machine learning for binary classification tasks. It contains structured information about passengers who were aboard the RMS Titanic during its ill-fated voyage in 1912. Each row in the dataset represents an individual passenger, while each column corresponds to a specific attribute describing personal, social, or economic characteristics. The primary objective associated with this dataset is to analyze these attributes to determine patterns related to passenger survival. The dataset includes a mixture of numerical and categorical features, providing a comprehensive representation of real-world data with varying types and distributions.

Dataset Parameters

Passenger Id : It is a unique identifier assigned to each passenger in the dataset. It is primarily used for indexing and distinguishing records and does not have any direct impact on analytical or predictive outcomes.

Survived : It is the target variable of the dataset, indicating the survival status of each passenger. A value of 1 represents that the passenger survived, while a value of 0 indicates that the passenger did not survive.

Pclass : It refers to the passenger class and represents the socio-economic status of individuals. It is categorized into three classes: first class (1), second class (2), and third class (3), with first class being the highest economic tier.

Name : The Name parameter contains the full name of the passenger r. It may also include titles such as Mr., Mrs., or Miss, which can provide additional contextual or demographic information.

Sex : It indicates the gender of the passenger and is categorized as male or female. It is an important attribute for understanding demographic distribution within the dataset.

Age : It represents the age of the passenger in years. This is a numerical feature and may include missing values in some records.

SibSp : It denotes the number of siblings or spouses traveling with the passenger aboard the Titanic. It provides insight into family or social connections.

Parch : It indicates the number of parents or children traveling with the passenger. Along with SibSp, it helps describe the family structure of each individual.

Ticket : It represents the ticket number assigned to each passenger. It is generally treated as a categorical attribute and **may** indicate group travel patterns.

Fare : It is a numerical feature that represents the amount paid by the passenger for the ticket. It often reflects the passenger's economic status and class.

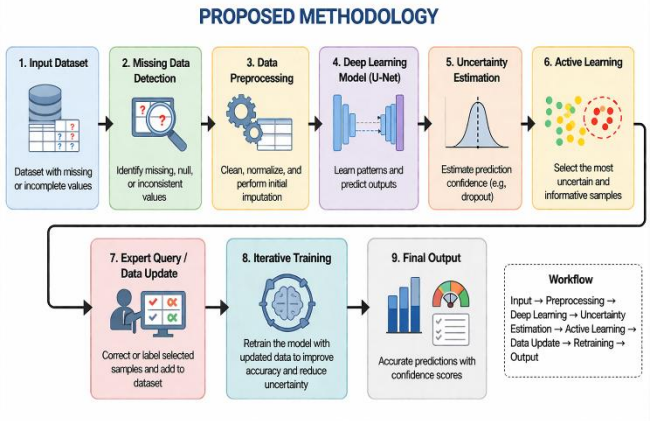
Cabin : It specifies the cabin number allocated to the passenger. This attribute contains many missing values and provides limited but potentially useful spatial information about accommodation.

Embarked : It indicates the port from which the passenger boarded the Titanic. It is a categorical feature with three possible values: Cherbourg (C), Queenstown (Q), and Southampton (S). give me the matter as dataset description for the base paper

IV. METHODOLOGY

The suggested system offers a deep learning framework that is aware of uncertainty and uses active learning to deal with datasets that have missing values. The workflow is designed to be done in steps and repeated to make predictions more accurate and reliable.

The process starts with the input dataset, which may have missing or incomplete values because of problems with collecting data in the real world. The system finds null, inconsistent, or undefined values across different features during the missing data detection stage.



Next, the data is preprocessed. This means that categorical features are encoded, numerical values are normalized, and initial imputation is done to make the dataset ready for model training. This step makes sure that the learning model gets clean and organized input.

Then, the processed data is put into a deep learning model that uses the U-Net architecture. This model can learn complex patterns and relationships even when the data is missing. The model makes predictions based on learned feature representations.

Estimating uncertainty is a big part of the system. This is where the model checks how sure it is about its predictions (for example, by using dropout techniques). This helps find samples where the model isn't sure what to do.

The active learning module chooses the most uncertain and useful data samples instead of using the whole dataset. This is where the expert query/data update stage comes in. The chosen samples are improved, fixed, or labeled before being added back to the dataset.

The system then goes through iterative training, which means that it uses the new dataset to retrain the model. This cycle of estimating uncertainty, choosing samples, updating data, and retraining goes on until the model gets better at its job and less uncertain.

Finally, the system gives you the output, which includes predictions and confidence scores. This makes the results easier to understand and more reliable.

V. RESULTS AND DISCUSSION

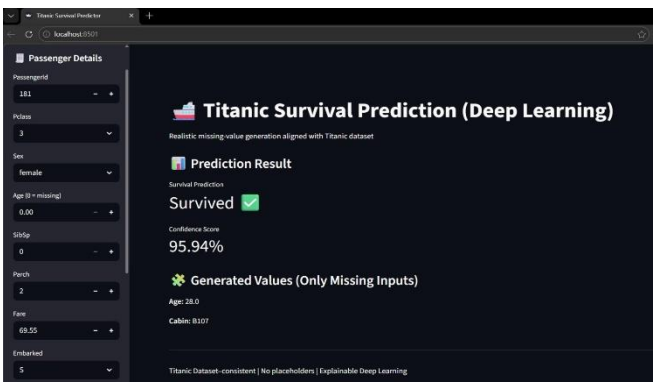


Figure 1: Web Application Interface for Passenger Input

The figure 1 shows the user interface of the deployed system, which was built with the Streamlit framework. There are fields on the interface where you can enter information about the passenger, such as their Passenger ID, class (Pclass), gender (Sex), age, number of siblings/spouses (SibSp), number of parents/children (Parch), fare, and embarkation point.

Each input field is well-organized, with dropdown menus for categorical features and boxes for continuous values. The system can handle incomplete inputs, so users can leave some fields blank or give them default values. This flexibility makes the app useful in real life, where you might not always have all the data you need. Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

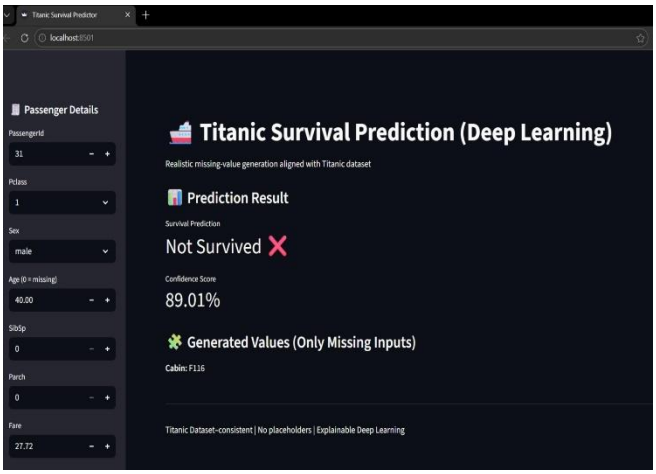


Figure 2: Result of the Prediction with Confidence Score

The figure shows what the system thought would happen based on the information about the passengers. The model tells you whether the passenger lived or died and shows you the result with a confidence score.

A high confidence score, like 95.94%, means that the model is very sure about its prediction. The system not only makes a prediction, but it also fills in missing information like age or cabin information based on patterns it has learned from data. This shows that the model can work with incomplete data and still give reliable results.

The experimental assessment validates the efficacy of the suggested uncertainty-aware deep learning framework in managing datasets with absent values. We trained the model on the preprocessed Titanic dataset and then used standard performance metrics like accuracy and loss to see how well it worked.

The model got better and better over time during training. At first, the loss was high and the accuracy was low, which meant that not much learning was happening. As training went on, the loss kept going down and the accuracy kept going up, which showed that the model was able to learn useful patterns from the data. The final model did well on test data that it had never seen before, which shows that it is good at generalizing and is strong.

The model's performance improved even more when uncertainty estimation and active learning were combined. The system made learning more efficient and less affected by

noisy or incomplete data by finding low-confidence predictions and focusing on informative samples. The iterative training process, in which uncertain samples are improved and added back to the dataset, made the system more accurate and reliable over time.

Using the Streamlit framework, the system was set up as an interactive web app that let users interact with it and make predictions in real time. The interface lets users enter information about passengers and automatically fills in missing values if they don't provide all of it.

The model makes a prediction about survival based on the inputs and gives a confidence score that shows how reliable it is. The findings indicate that the suggested framework yields precise and reliable predictions, even with incomplete data, rendering it appropriate for practical application.

VI. CONCLUSION AND FUTURE WORK

Conclusion:

This study introduces an uncertainty-aware deep learning framework for addressing missing data, utilizing a U-Net architecture combined with active learning. Traditional imputation methods frequently do not account for intricate data relationships, resulting in erroneous predictions. To fix this, the suggested method treats missing data as a reconstruction problem, which lets the model learn useful patterns and guess missing values correctly.

The U-Net architecture is good at taking features out and putting them back together, and masked learning makes sure that only missing values are changed and not existing data. Adding uncertainty estimation to the model lets it figure out how sure it is about its predictions and find outputs that aren't as reliable. Active learning makes things even more efficient by picking the samples that are both the most informative and the most uncertain for training.

Experimental results show that the suggested method makes data better, predictions more accurate, and models more general. The framework is a dependable and effective way to deal with incomplete datasets in real-world situations.

Future Work:

The U-Net architecture with attention mechanisms is the main focus of future improvements to the proposed uncertainty-aware deep learning framework for active learning with missing data. The model can focus on the most important features during reconstruction by adding attention layers. This makes the model more interpretable and helps improve the accuracy of missing value imputation. Also, making the framework work with time-series and sequential data by using models like LSTM or GRU can make it useful for real-world uses like healthcare monitoring, IoT sensor systems, and financial analysis, where values can be missing over time.

Another important goal is to make the framework's uncertainty estimation and scalability better. Bayesian deep learning and ensemble models are two advanced probabilistic methods that can give reconstructed values more reliable confidence scores. This makes the system more reliable in important areas. You can also use VAE or GAN-based hybrid imputation methods with the framework to make missing data reconstruction even stronger. Also, making the system better for big datasets, faster training, and deployment on the cloud or the web can make it easier to use in real life and let the model work well in many different areas.

REFERENCES

[1] A. Tharwat and W. Schenck, “Balancing exploration and exploitation: A novel active learner for imbalanced data,” Knowl.-Based Syst., vol. 210, Dec. 2020, Art. no. 106500.

[2] W. Liu, H. Zhang, Z. Ding, Q. Liu, and C. Zhu, “A comprehensive active learning method for multiclass imbalanced data streams with concept drift,” Knowl.-Based Syst., vol. 215, Mar. 2021, Art. no. 106778.

[3] A. Tharwat and W. Schenck, “A conceptual and practical comparison of PSO-style optimization algorithms,” Expert Syst. Appl., vol. 167, Apr. 2021, Art. no. 114430.

[4] A. Tharwat, “Classification assessment methods,” Appl. Comput. Infor mat., vol. 17, no. 1, pp. 168–192, Jan. 2021.

[5] C. K. Enders, “Applied Missing Data Analysis”. New York, NY, USA: Guilford, 2022.

[6] A. Tharwat and W. Schenck, “A novel low-query-budget active learner with pseudolabels for imbalanced data,” Mathematics, vol. 10, no. 7, p. 1068, Mar. 2022.

[7] Z. Chai, C. Zhao, B. Huang, and H. Chen, “A deep probabilistic transfer learning framework for soft sensor modeling with missing data,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 7598–7609, Dec. 2022.

[8] M. P. likowski, M. Śmieja, Ł. Struski and J. Tabor, "MisConv: Convolutional Neural Networks for Missing Data," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 2917-2926, doi: 10.1109/WACV51458.2022.00297.

[9] A. Tharwat and W. Schenck, "Active Learning for Handling Missing Data," in IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 2, pp. 3273- 3287, Feb. 2025, doi: 10.1109/TNNLS.2024.3352279.