University of Ibadan
Journal of Science and Logics in ICT Research

A Predictive Model for Uncertainty Analysis on Big Data Using Bayesian Convolutional Neural Network (CNN)

[1] Obasi, E. C. M. and [2] Stow, M. T.

Department of Computer Science and Informatics, Federal University Otuoke, Nigeria

[1]anchinos@yahoo.co.uk, [2]stowmt@fuotuoke.edu.ng

Abstract

The need of addressing uncertainty in big data increases as more data is created and examined. It is essential to comprehend, measure, and control uncertainty in large data for dependable and useful analysis. Uncertainty in big data analysis is one of the major problems of big data, and if not handled correctly, it will lead to wrong predictions/classification of the model. In order to solve the problem of uncertainty in big data, this paper presents a Bayesian CNN model for the prediction of uncertainty in big data. The Bayesian CNN model uses a probability score in predicting uncertainties in big data. With this, it does not just show the classified results that were made by the model, it also shows the probability score, which signifies the decision score of the model when making classifications on images. The result of Bayesian model shows a better result of 99.9% for both training and testing.

Keywords: *Big Data Analysis, Uncertainty, Bayesian CNN, Chest X-ray*

1. INTRODUCTION

Big data analytics has attracted great attention from both academia and industry as the desire to gain more knowledge in trends of stupendous datasets is the increase. As the amount, variety, and speed of data increases, so too does the uncertainty inherent within, leading to a lack of confidence in the resulting analytics process and decisions made thereof. Large amounts of data provide a challenge for analysts because of the inherent uncertainty introduced by factors including measurement error, missing data and variation in data quality. The need of addressing uncertainty in big data increases as more data is created and examined. This prompts the data driven optimization that can integrate machine learning for decision making under uncertainty and identifies potential research opportunity in the business field of Bayesian optimization. It is essential to comprehend, measure and control uncertainty in large data for useful analysis [1].

1.1. Means of Dealing with Uncertainty in Large Data set.

Applying probabilistic models is one strategy for dealing with uncertainty in large data. Since uncertainty can be included in probabilistic models, they are often more accurate and resilient in their predictions. For example, Probabilistic models give a systematic method for addressing uncertainty in data and have been utilised effectively in a broad variety of applications, from natural language processing to computer vision [2].

Using ensemble algorithms is another strategy for handling uncertainty in huge data. When applied to the analysis, the use of several algorithms to the models' predictions (as is done by ensemble techniques) increases both accuracy and resilience. To mitigate the effects of data variability and uncertainty, "ensemble approaches have become more popular for large data analysis. The sheer amount and complexity of the data are one of the difficulties in dealing with uncertainty in big data. As data volumes increase, the computational complexity of probabilistic models and ensemble approaches may become intractable [3]

As if the technical difficulties of massive data analysis weren't enough, there are also the statistical difficulties of dealing with uncertainty. High-dimensional and non-linear data structures are ubiquitous in big data

Obasi E.C.M. and Stow M. T. (2023). A Predictive Model for Uncertainty Analysis on Big Data Using Bayesian CNN. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 9 No. 1, pp. 52 – 62.

research, and conventional statistical approaches may struggle to deal with them. New statistical approaches are required to cope with the problems of large data. These issues include "the curse of dimensionality, sparsity, and nonlinearity. Using machine learning methods, such as deep learning, is one way to manage uncertainty with massive data. When it comes to reducing uncertainty and increasing the reliability of predictions, deep learning algorithms' ability to learn complicated representations of the data is invaluable. Deep learning has demonstrated encouraging results in a broad variety of applications, from voice recognition to picture classification, and has the potential to overcome many of the issues of large data analysis [4].

Data fusion methods are another option for dealing with large data's inherent ambiguity. Using data fusion, researchers may pool information from many databases to improve an analysis's validity and reliability. Data fusion has been utilized effectively in various applications, from military surveillance to medical diagnosis, and may assist to limit the effects of ambiguity in the data. Data fusion in big data analysis presents a number of difficulties, one of which is the necessity to deal with data from a number of different modalities and sources. Because of the data's inherent diversity, efficient amalgamation is often a challenge. Data fusion is difficult in big data analysis because of the requirement to combine information from several sources of varied quality and trustworthiness [5].

The main aim of uncertainty prediction in big data analysis is to improve the quality of statistical models and forecasts. Uncertainty can originate from a number of sources in big data analysis, including but not limited to noisy or missing data, model assumptions, or the limitations of research methodologies.

The ability to foresee uncertainty allows decision-makers to evaluate the validity of the analysis and make better choices. It can also point out places where more data gathering, cleaning, or pre-processing is needed because of possible noise or inconsistencies. The justification of the study includes:

1. Robust Decision Making: In many big data analysis applications, decisions are made based on statistical models and predictions. Predicting uncertainty can help decision-makers to assess the reliability of these models and make more robust decisions.
2. Risk Management: Predicting uncertainty can help mitigate risks associated with big data analysis, such as incorrect or misleading conclusions, by providing insight into the accuracy and reliability of the analysis.
3. Optimization of Data Analysis: Predicting uncertainty can help optimize big data analysis by identifying areas where the data may be noisy, inconsistent, or incomplete. This can help improve the quality of the data analysis and lead to more accurate and reliable results.
4. Better Resource Allocation: Predicting uncertainty can help allocate resources more effectively by identifying areas where additional data collection, cleaning, or pre-processing is needed. This can help optimize the use of computational resources and reduce the time and cost of data analysis.

2. RELATED WORKS

Machine learning algorithms, geographic information systems, and physical models were all used to determine the technical photovoltaic (PV) system.

The potential of specific roof areas on an hourly basis by Walch et al. [6] assessed the uncertainties associated with each stage of the potential evaluation, combining those estimates to get a quantitative value for the uncertainty in the final PV potential. The methodology is tested on 9.6 million roofs in Switzerland and may be applied to any big area or nation with enough data. Future energy systems with decentralized electrical grids might benefit from the described approach for hourly rooftop PV potential and uncertainty estimates. The findings may be utilized to develop workable policies for installing solar panels on roofs.

Gholizadeh et al. [7] created a hybrid model to forecast the outcome of events with a high

degree of uncertainty. With the use of big data, the model is a mixed-integer nonlinear software that may help one make the most responsible choices for sustainable purchasing and shipping. The massive data issue is tackled using a heuristic technique that employs a powerful fuzzy stochastic programming strategy. In order to forestall disruptions, the suggested model employs a stochastic programming strategy that is scenario-based.

Ning and You [8] provides promising leads for further study by focusing on scenario-based optimization using the strength of deep learning and a closed-loop data-driven optimization framework that permits input from mathematical programming to machine learning. Perspectives on data-driven multistage optimization with online learning and an iterative learning-while-optimizing approach are discussed.

Shukla and Muhuri [9] provide a solution to the unpredictability of Big Data, the authors recommend using fuzzy clustering. Since genes may be part of several clusters using fuzzy clustering, they can be involved in more than one kind of cellular function, as well as subcellular variations and metabolic pathways. Different cluster validity metrics have been used to examine the impact of the generated uncertainty on large data clustering. Results from clustering using IT2 fuzzy uncertainty modeling are compared to those using type-1 fuzzy sets-based uncertainty modeling. Their results show that the proposed IT2-FS-based method can effectively cluster huge gene expression datasets with high degrees of uncertainty and improve upon previous methods.

Fahmideh and Beydoun [10] uses a model centered on desired outcomes to pinpoint the causes of quality-related project failure and the best ways to fix them. This method uses a combination of fuzzy logic and exploratory data analysis to identify the best set of architectural options for enabling industrial systems with big data. The proposed method improves upon the current state of the art in two ways: (i) by providing a goal-oriented model for exploring requirements and barriers to integrating manufacturing systems with big data analytics platforms, and (ii) by providing a systematic analysis of the architectural decisions under uncertainty that takes into account the preferences of stakeholders.

The subject of Bayesian optimization under uncertainty through the perspective of current data by Wang and He [3] highlights significant research problems and the promise of data-driven optimization that naturally incorporates fuzzy, machine learning, and deep learning for decision-making under uncertainty. Science and database data mining face significant challenges due to Big Data. Here, the authors took a peek at the intriguing things my community has been up to at this conference in order to address the big data issue.

Complete Review of Uncertainty Prediction in Big Data, by E. Tuncay, S. Aydin, and M. Isik [11], Methods for predicting uncertainty in large datasets are discussed in this study. Topics covered include Bayesian networks, machine learning, and probabilistic modeling. The obstacles and potential developments in this area are also addressed by the authors.

Zhang *et al.* [12] present a review of the literature on uncertainty quantification techniques for use in big data analytics. Probabilistic graphical models, Monte Carlo approaches, and sensitivity analysis are only some of the tools discussed in this study for quantifying uncertainty in big data analytics. The writers also include case examples to show how these techniques might be used in the real world.

The paper "Deep Learning for Uncertainty Prediction in Big Data," written by Mishra and Ganguly [13], provides an in-depth analysis of the available techniques for making predictions about the future using deep learning and large datasets. Several deep learning architectures, such as convolutional neural networks, recurrent neural networks, and generative models, are discussed, along with their uses in uncertainty prediction.

A Review of Uncertainty Prediction Methods for Big Data, by Venkataraman and Jayaraman [14] surveys some of the most common approaches to making predictions in the face of uncertainty when working with large datasets, such as Bayesian inference, Markov chain Monte Carlo, and probabilistic modeling. Both

the opportunities and the threats that exist in this field are addressed by the authors.

"Uncertainty Quantification and Prediction in Big Data Analytics: A Survey" is a paper written by Zhou *et al.* [15], which adopted Bayesian inference, Monte Carlo methods, and machine learning strategies are only some of the uncertainty quantification and prediction techniques. The authors provide a comprehensive overview of their application to big data analytics. The writers also talk about the new developments and where they think the discipline should go in terms of research.

3. METHODOLOGY

This session describes the system architecture and the various components that are made up of the system architecture in predicting and analyzing uncertainty in big data. A detailed design of the proposed system architecture can be seen in Figure 1.

Dataset: A chest X-ray data is used here. The dataset comprises 5,863 Chest X-ray images which can further be divided into 3 main folders (train, test, val) with further subfolders for each picture type (Pneumonia/Normal). The dataset was downloaded from Kaggle.com.

Retrospective cohorts of paediatric patients aged one to five years old were chosen from the Guangzhou Women and Children's Medical Center's chest X-ray imaging database (Kaggle). The X-rays of the patient's chest were taken as part of their regular medical treatment.

Before any chest x-ray pictures could be analysed, they all had to pass a quality control screening, which meant that any scans that were too blurry or otherwise unusable were thrown out. Before the AI system could be trained, the picture diagnoses were reviewed and scored by two doctors. The assessment set was double-checked by a third specialist to account for any grade discrepancies.

Data Pre-processing: The pre-processing of the dataset (chest x-ray images) has to do with the conversion of the dataset labels into binary numbers (0s and 1s). This was achieved using the image generator function in reading the number of images (Chest X-ray) in the folder, and assigning 0s to images in the folder with the name Normal and 1s to the folder with the name Pneumonia.

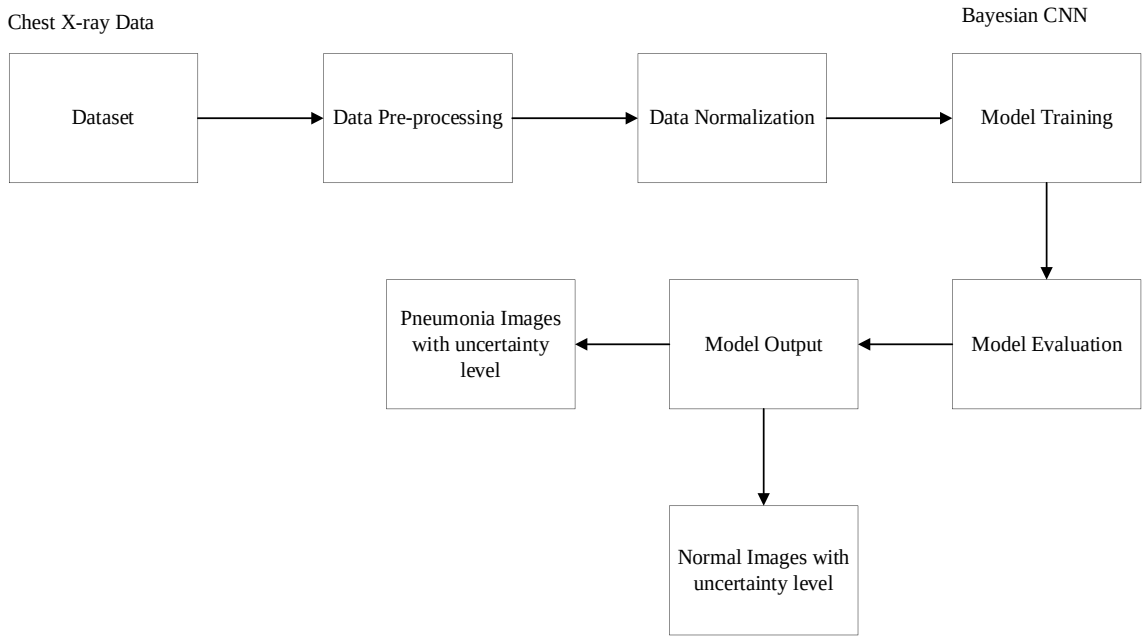

Figure 1: Architectural Design

Image Normalization: This has to do with the resizing and scaling of the images before using them as input for the training of the Bayesian neural network model. This was achieved by using the image. resize function in python. This can be represented as image. resize (225,225).
Bayesian CNN: Bayesian CNNs can be used to predict uncertainty in big data by modelling the posterior distribution of the model parameters. The posterior distribution represents the updated beliefs of the model after it has been trained on the data. The uncertainty can be estimated by calculating the variance of the posterior distribution of the predicted outputs.

Let X and Y represent the Input and Output of our data, The Bayesian CNN to predict the output Y given the input X. We can use the following mathematical expressions to model the posterior distribution of the model parameters:

1. Prior distribution: P(θ) represents the prior distribution of the model parameters θ. The prior distribution represents our initial beliefs about the distribution of the parameters.
2. Likelihood function: P(Y|X,θ) represents the likelihood function, which models the probability of the output Y given the input X and the model parameters θ.
3. Posterior distribution: P(θ|X,Y) represents the posterior distribution of the model parameters θ, given the input X and the observed output Y. The posterior distribution is proportional to the product of the prior distribution and the likelihood function, as shown in Bayes' rule: P(θ|X,Y) = P(Y|X,θ) * P(θ) / P(Y|X).
4. Predicted output: Once we have the posterior distribution of the model parameters, we can use it to make predictions on new data. The predicted output is the expected value of the posterior distribution, given the input X: Y_pred = E[P(Y|X, θ)] = $\int$ Y * P(Y|X, θ) dY
5. Uncertainty estimation: The uncertainty can be estimated by calculating the variance of the posterior distribution of the predicted outputs: Var(Y_pred) = E [(Y_pred - E[Y_pred]) ^2] where E[Y_pred] is the expected value of the predicted output, as calculated in step 4.

   By modeling the posterior distribution of the model parameters, we can estimate the uncertainty in the predicted outputs, which is important for making informed decisions in big data applications.

Model Evaluation: Accuracy, precision, recall, and F1-Score were utilized in an assessment matrix to measure the effectiveness of a random forest model. Specifically, we employed a classification report and a confusion matrix to explain the results of the model. This was used to explain why the model was sometimes correctly predicted and sometimes incorrectly predicted the test data. True positive, true negative, false positive, and negative denote the accurate and incorrect predictions, respectively.

4. Results and Discussion

This paper presents a predictive model for uncertainty analysis on Big Data. The first phase has to do with explorative data analysis and the second phase has to do with the Bayesian CNN network for uncertainty.

4.1 Phase 1: Exploratory Data Analysis (EDA).

We conducted EDA, so that we can have a proper view of the data, and the trend of the dataset. From the EDA, we can have a visualized view of chest x-ray images. This can be seen in Figure 2. Figure 3 also shows the count of the training data and Figure 4 shows a countplot for the test data.

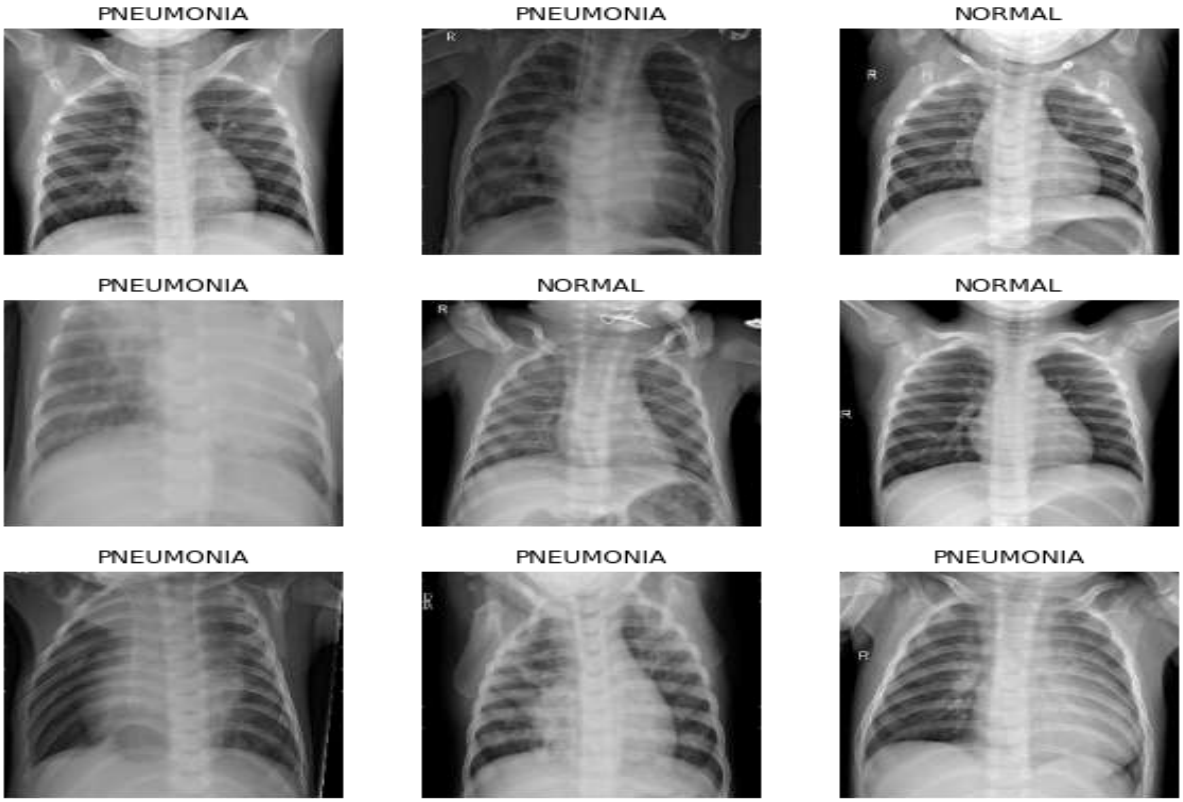

Figure 2. Visualized Result of the Chest X-ray Images

Figure 2 shows the scanned chest x-ray images of the dataset. The visualized images comprise of both normal and pneumonia images.

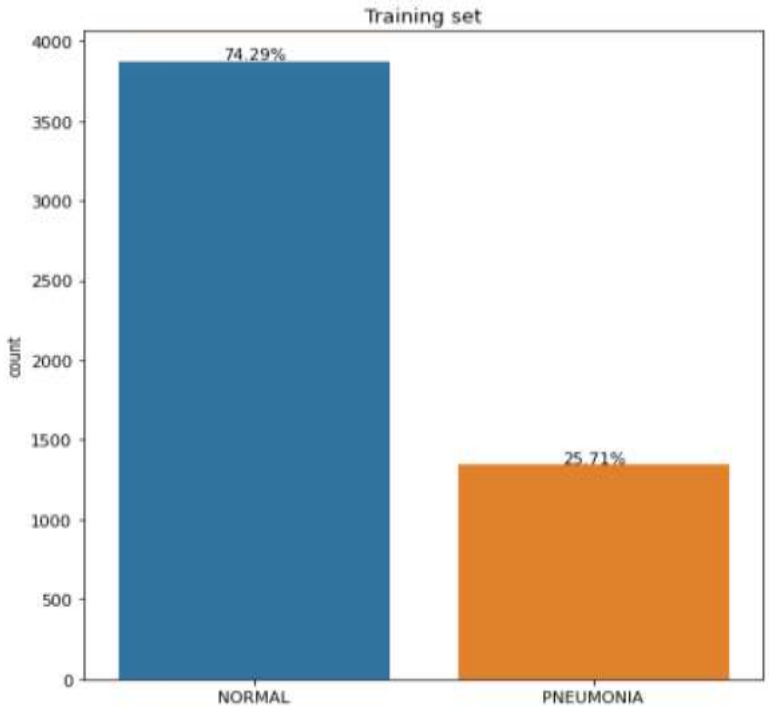

Figure 3: Histogram of the Training data

Figure 3 shows a count plot of the number of classes and their number of occurrences. Here the normal images appear at about 3800, and the pneumonia images appear at about 1400.

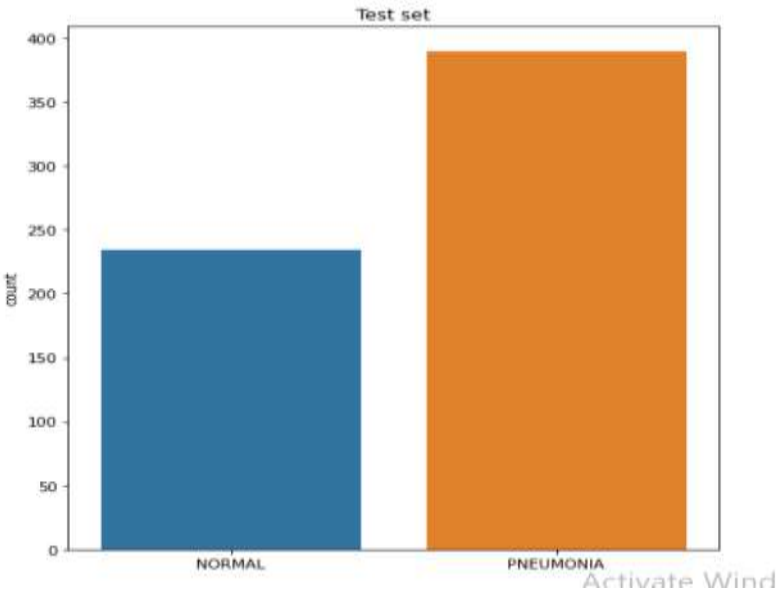

Figure 4: Histogram of the Test data

The histogram here shows a count plot of the number of classes and their number of occurrences. Here the normal images appear at about 240, and the pneumonia images appear at about 380.

4.2: Phase 2: Model Training with Bayesian CNN

The model was trained using Bayesian Convolutional Neural Network. The Bayesian model was trained using 3 layers. The first layer contains an input shape=(224,224,3), with weight=''imagenet. The second layer contains the following parameters:

filters = 64,
kernel_size = 3,
padding = 'same',
activation = tf.nn.silu,
pool_size = (2, 2),
strides = (1, 1),
name = 'residual_block1

The Third layer contains the following parameters:

filters = 128
kernel_size = 3,
padding = 'same',
activation = tf.nn.silu,
pool_size = (2, 2),
strides = (1, 1),
name = 'residual_block1

Other hyperparameters used in training the model are loss= categorical, optimizer=adam, epoch=140, and batch_size=128. The training result displays the mean squared error for both the training and validation test. Figures 5, and 6 show the graphical analysis of the model's performance using accuracy and loss for 140 training steps. Figures 7 and 8 show the predicted result of the Bayesian CNN model.

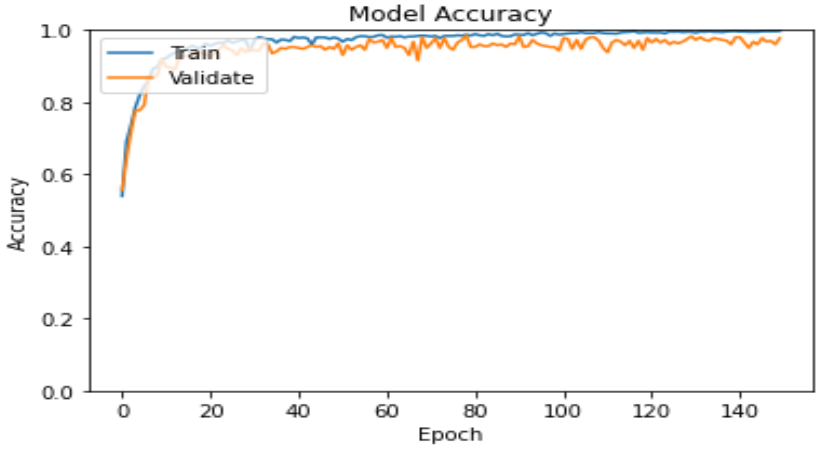

Figure 5: Accuracy for both Training and Testing Data.

Figure 5 shows the accuracy of the Bayesian model for predicting uncertainty. The result shows that the model achieved both training and validation accuracy of 99%.

Figure 6 shows the accuracy of the Bayesian CNN model for the prediction of uncertainty in big data. The result shows that the model had both training and validation loss below 0.01%.

In Figure 7, the model classified the result to be of the normal category. The probability of the model is used to check the level of uncertainty. Here, we can see that the model struggled to make the right classification with a probability level of 1.0 for normal and a probability score of 0.63 for the pneumonia images.

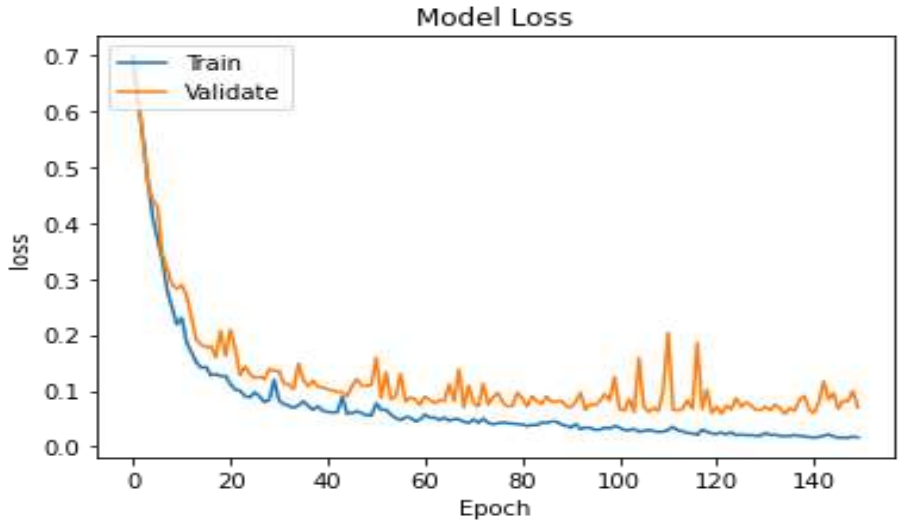

Figure 6: Loss values for both Training data and Testing Data

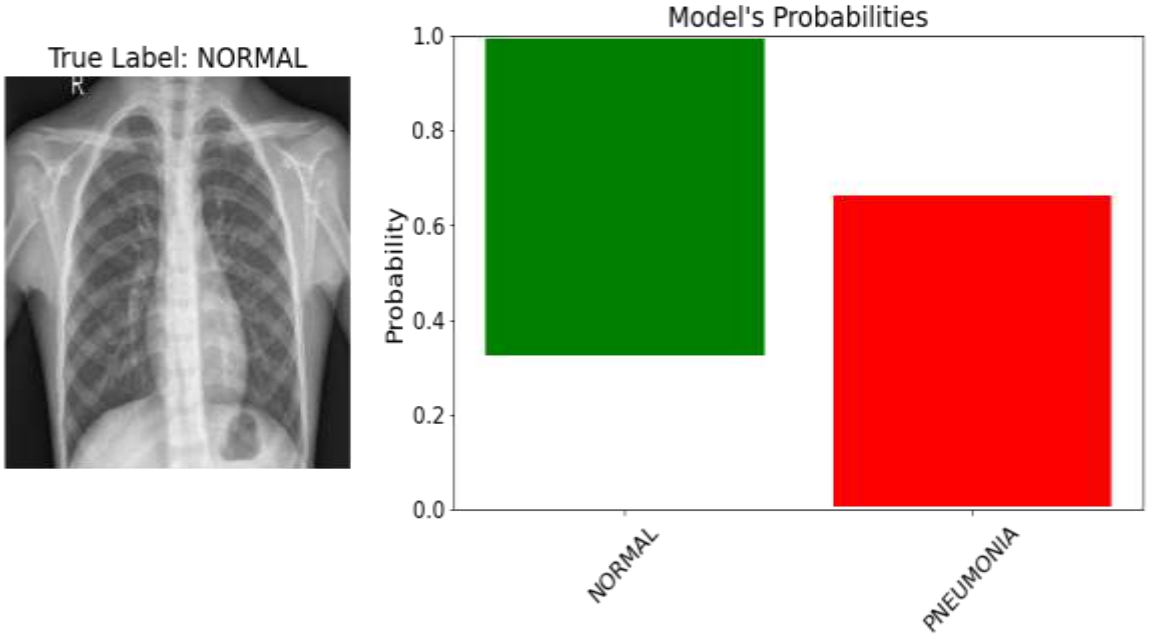

Figure 7: Predicted Result Normal

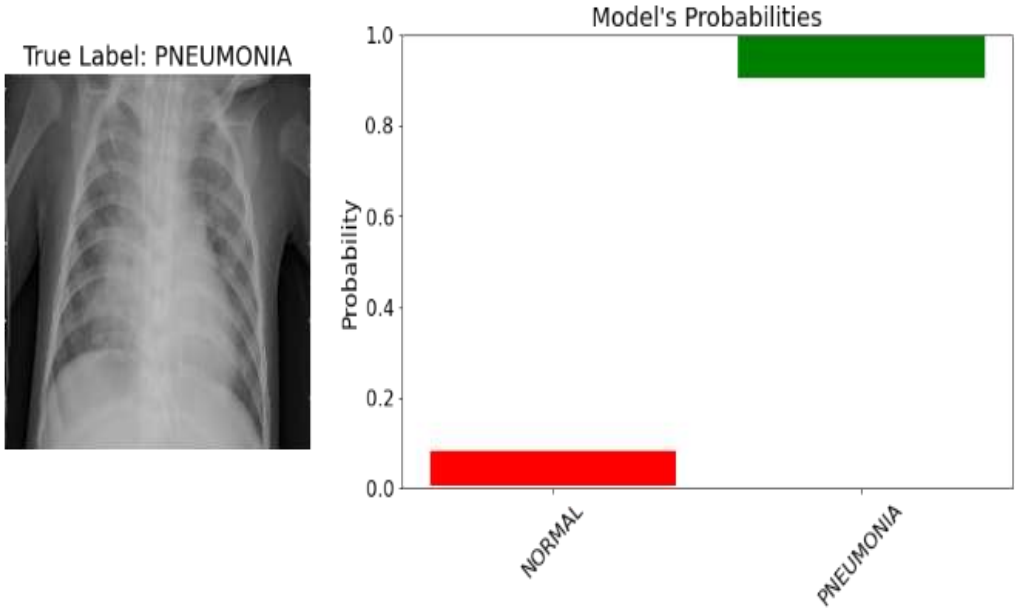

Figure 8: Predicted Result Pneumonia

In Figure 8, the model classified the result to be of the pneumonia category. The probability of the model is used to check the level of uncertainty. Here, we can see that the model did not struggle to make the right classification with a probability level of 0.10 for normal and a probability score of 1. for the pneumonia images.

4.3 Comparison of results with other existing systems

This session describes the comparison of our proposed system (Bayesian CNN) with other existing systems. Table 1 shows an existing system, the technique used, and the evaluation of their techniques in terms of accuracy.

Table 1: Model Comparison with other existing Systems

Authors	Title	Accuracy Score (%)
Fadi *et al.* [16]	Quantifying Uncertainty in Internet of Medical Things and Big-Data Services Using Intelligence and Deep Learning	95
Chen et al. [17]	A novel fuzzy deep-learning approach to traffic flow prediction with uncertain spatial–temporal data features	97
Our Method	A Predictive Model For Uncertainty Analysis On Big Data Using Bayesian Cnn	99

4.4 Discussion

From the experiment conducted, Figure 2 shows the Chest X-ray images. The visualized images comprise 6 pneumonia images and 3 normal images. Figure 3 shows both the number of pneumonia and normal images that are present in the training data set. The histogram shows that 3800 images are pneumonia and 1400 are normal images. Figure 4 also shows both the number of images that of class pneumonia and normal. The histogram here shows a count plot of the number of classes and their number of occurrences. Here the normal images appear at about 240, and the pneumonia images appear at about 380.

Figure 5 shows the accuracy obtained by the Bayesian CNN model during training and testing at each training step. The blue line represents the training score and the orange line represents the testing score. Here the Bayesian model scored 99% for both training and testing scores. Figure 6 shows the loss values obtained by the Bayesian CNN model during training and testing at each training step. The blue line represents the training loss and the orange line represents the testing loss. Here the Bayesian model got below 0.01% for both training and testing data.

Figure 7 shows the predicted result of the Bayesian CNN model and the probabilistic score. Here, the model classified the result to be of the normal category. The probability of the model is used to check the level of uncertainty. Here, we can see that the model struggled to make the right classification with a probability level of 1.0 for normal and a probability score of 0.63 for the pneumonia images.

Figure 8 shows the predicted result of the Bayesian CNN model and the probabilistic score. Here, the model classified the result to be of the pneumonia category. The probability of the model is used to check the level of uncertainty. Here, we can see that the model did not struggle to make the right classification with a probability level of 0.10 for normal and a probability score of 1 for the pneumonia images.

5. CONCLUSION

Uncertainty in big data analysis is one of the major problems of big data, and if not handled correctly, it will lead to wrong predictions/classification of the model. In order to solve the problem of uncertainty in big data, this paper presents a Bayesian CNN model for the prediction of uncertainty in big data. The Bayesian CNN model uses a probability score in predicting uncertainties in big data. With this, it does not just show the classified results that were made by the model, it also shows the probability score from 0 to 1, which signifies the decision score of the model when making classifications on images.

The result of the Bayesian model shows a better result of about 99.9% for both training and testing. The Bayesian CNN model was also compared with other state of art models. The state of art models had 95%, and 97%, while our method had 99% in predicting uncertainty in Big data. This shows that our model is more outstanding than the state of art models.

References

[1] Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data, 6(1), 1-16.

[2] Antonopoulou, H., Mamalougou, V., & Theodorakopoulos, L. (2022). The role of economic policy uncertainty in predicting stock return volatility in the banking industry: A big data analysis. Emerging Science Journal, 6(3), 569-577.

[3] Wang, X., & He, Y. (2016). Learning from uncertainty for big data: future analytical challenges and strategies. IEEE Systems, Man, and Cybernetics Magazine, 2(2), 26-31.

[4] Millie, D. F., Weckman, G. R., Young II, W. A., Ivey, J. E., Fries, D. P., Ardjmand, E., & Fahnenstiel, G. L. (2013). Coastal 'Big Data'and nature-inspired computation: Prediction potentials, uncertainties, and knowledge derivation of neural networks for an algal metric. Estuarine, Coastal and Shelf Science, 125, 57-67.

[5] Afshari, H., & Peng, Q. (2015). Using big data to minimize uncertainty effects in adaptable product design. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 57113, p. V004T05A052). American Society of Mechanical Engineers.

[6] Walch, A., Castello, R., Mohajeri, N., & Scartezzini, J. L. (2020). Big data mining for the estimation of hourly rooftop photovoltaic potential and its uncertainty. Applied Energy, 262, 114404.

[7] Gholizadeh, H., Fazlollahtabar, H., & Khalilzadeh, M. (2020). A robust fuzzy stochastic programming for sustainable procurement and logistics under hybrid uncertainty using big data. Journal of Cleaner Production, 258, 120640.

[8] Ning, C., & You, F. (2019). Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming. Computers & Chemical Engineering, 125, 434-448.
[9] Shukla, A. K., & Muhuri, P. K. (2019). Big-data clustering with interval type-2 fuzzy uncertainty modeling in gene expression datasets. Engineering Applications of Artificial Intelligence, 77, 268-282.
[10] Fahmideh, M., & Beydoun, G. (2019). Big data analytics architecture design—An application in manufacturing systems. Computers & Industrial Engineering, 128, 948-963.
[11] Tuncay, E., Aydin, S., & Isik, M. (2018). Uncertainty prediction in big data: A comprehensive review. Journal of Big Data, 5(1), 1-24.
[12] Zhang, J., Liu, J., & Guo, S. (2020). A review of uncertainty quantification methods in big data analytics. IEEE Access, 8, 75206-75220.
[13] Mishra, A., & Ganguly, A. R. (2021). Deep learning for uncertainty prediction in big data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(1), e1406.
[14] Venkataraman, S., & Jayaraman, K. R. (2019). A survey of uncertainty prediction techniques in big data. Journal of Big Data, 6(1), 1-20.
[15] Zhou, Y., Liu, H., & Chen, C. H. (2018). Uncertainty quantification and prediction in big data analytics: A survey. Journal of Big Data, 5(1), 1-29.
[16] Al-Turjman, F., Zahmatkesh, H., & Mostarda, L. (2019). Quantifying uncertainty in internet of medical things and big-data services using intelligence and deep learning. IEEE Access, 7, 115749-115759.
[17] Chen, W., An, J., Li, R., Fu, L., Xie, G., Bhuiyan, M. Z. A., & Li, K. (2018). A novel fuzzy deep-learning approach to traffic flow prediction with uncertain spatial–temporal data features. Future generation computer systems, 89, 78-88.