Probabilistic Deep Learning for Energy Time Series Forecasting: A Comparative Study

Sunil Pradhan Sharma*, **Elakkiya Daivam****
*Senior Lead Software Engineer, Capital One Services, LLC
**Lead Software Engineer, Capital One Services, LLC

Abstract— **Deep learning methods have achieved widespread success and are continually being utilized in new domains, including the prediction of energy time series. However, these methods do not indicate their confidence in predictions, hindering the ability to evaluate, interpret, and enhance forecast accuracy. As a result, this research delves into the exploration of probabilistic extensions to deep learning and their implementation in energy time series forecasting. The evaluated methods include concrete dropout, deep ensembles, Bayesian neural networks, deep Gaussian processes, and functional neural processes. Two load forecasting scenarios are considered for evaluation: short-term (single-step) and day-ahead (multi-step) forecasting. The methods are assessed based on calibration, sharpness, and proficiency in indicating a lack of knowledge. To provide a reference, a simple neural network and a quantile regression model are also employed. Overall, the methods demonstrate strong performance, with concrete dropout, deep ensembles, and Bayesian neural networks performing equally well or better than the reference methods. Functional neural processes and deep Gaussian processes exhibit weaker performance, likely attributed to convergence issues and suboptimal parameters. Notably, deep ensembles prove to be straightforward to implement and train, requiring very few hyperparameters.**

Index Terms— **Day-Ahead, Deep Learning, Energy Time Series, Short-Term**

I. Introduction

The recent decade has witnessed significant advancements in neural networks and deep learning, attributed to the increased processing power that has enabled the training of complex models within reasonable time frames. This has led to the widespread application of deep learning methods across various fields, particularly in the notable success of image recognition with Convolutional Neural Networks (CNNs) [1], [2], [3], [4], [5], [6]. In theory, a neural network can approximate any function given a sufficient number of neurons. While other methods, such as Support Vector Machines (SVMs), possess similar theoretical function approximation capabilities, the inherent strength of neural networks lies in their ability to be layered into a deep structure, allowing for the extraction of intrinsic feature representations during training. However, it is important to note that these representations particularly have flaws, as they are heavily reliant on the training data and can sometimes prove to be counterintuitive. The success of neural networks has spurred research into the development of deep models that integrate various methods, such as incorporating neural networks with other domains like graphs or latent variables to create novel methods, along with efforts to address the limitations of neural networks themselves through techniques like generating adversarial training data and exploring probabilistic extensions.

Despite their broad application and success, standard neural networks do have a significant limitation; they cannot provide an estimate of their confidence in their predictions. The absence of uncertain information places users at a disadvantage by not allowing them to fully understand the model's learning and areas for potential improvement and limiting their ability to make informed decisions based on the model's output. In many fields, having an estimate of prediction confidence is crucial for interpreting and utilizing the results effectively. For instance, a low-confidence prediction may require further scrutiny rather than being used as is. The primary objectives of this study are to explore probabilistic extensions to deep learning, apply them in the domain of energy time series forecasting, and evaluate the results. To achieve this, various probabilistic extensions to neural networks, such as Bayesian neural networks, deep ensembles, functional neural processes, concrete dropout, and deep Gaussian processes, are examined. At present, there is limited research in the area of probabilistic deep learning for forecasting, hence configuring and evaluating probabilistic deep models in the energy time series domain presents an opportunity to contribute to scientific progress, establish deep learning approaches as standard methods, and bridge the gap in uncertainty estimation compared to statistical forecasting approaches. Given the volatility introduced by the continuing growth of renewable energy, particularly solar and wind power, a confidence estimate for forecasts is essential for making more informed decisions amidst increasing uncertainty in the energy system. The methods are evaluated in a load forecasting scenario and compared to a simple neural network and a quantile regression model as reference. The findings demonstrate that concrete dropout and deep ensembles offer significant advantages in terms of simplicity and training time, while functional neural processes and deep Gaussian processes tend to converge at a slower rate and did not perform as well overall, potentially indicating the need for further optimization. Notably, these models also base their confidence estimates not only on the distribution of the training data but also on their confidence in the similarity of a situation to the relevant training data.

The study is as follows; similar papers are shown in the following section. The materials and methods are provided in Section III. The experimental analysis is carried out in Section IV, and in Section V, we provide some conclusions and plans for future research.

II. Related Works

The rise of processing power in the last decade has facilitated the rapid advancement of deep learning techniques, particularly in fields like image recognition [7], [8], where CNNs have seen remarkable success. Unlike other methods such as SVMs [9], [10], [11], deep neural networks [12], [13] can extract complex feature representations through layered structures, albeit with some reliance on training data and potential counterintuitive behaviour. The success of neural networks has spurred the development of deep learning models integrating various domains and addressing their limitations. However, a significant limitation of standard neural networks is their inability to estimate confidence in predictions according to [14], which is essential for informed decision-making. To address this, we aim to explore and evaluate various probabilistic extensions such as Bayesian neural networks, deep ensembles, functional neural processes, concrete dropout, and deep Gaussian processes in the context of energy time series forecasting. Given the limited research in probabilistic deep learning for forecasting such as [15], [16], [17], [18], our study aims to contribute to scientific progress, establish deep learning approaches as standard methods, and bridge the gap in uncertainty estimation compared to statistical forecasting approaches. With the increasing uncertainty in the energy system due to the growing renewable energy sector, confidence estimates in forecasts are crucial for informed decision-making. In our load forecasting scenario, we found that concrete dropout and deep ensembles offer significant advantages in simplicity and training time while indicating the need for further optimization in the case of functional neural processes and deep Gaussian processes. Moreover, the introduced models demonstrated similar performance to the reference models while providing confidence estimates based on the distribution of training data and their confidence in the similarity of a situation to the relevant training data. In a related study by [19], deep learning was employed in a probabilistic manner for wind power forecasting, demonstrating the growing interest in this area. They applied wavelet decomposition and CNNs to the data, showcasing the applicability of deep learning in energy forecasting [20], [21], [22], [23], [24], [25].

III. Materials and Methods

Not all introduced methods are utilized for evaluation purposes. Mixture density networks and quantile regression for neural networks are excluded from consideration as they solely provide an estimate of the aleatoric uncertainty. Both methods can be applied to the other models and therefore will not be implemented on their own. The methods chosen for further use include concrete dropout, deep ensembles, Bayesian neural networks, deep Gaussian processes, and functional neural processes. Monte-Carlo dropout is a straightforward and mathematically validated method that is widely used in related research. Concrete dropout is selected as a more advanced iteration based on the principles of Monte-Carlo dropout. Similarly, deep ensembles are chosen as a straightforward approach. The decision was made against implementing the adversarial training extension from the paper as it might potentially enhance any of our methods. Furthermore, an implementation of Bayesian neural networks applying the concept of distributions over weights and biases directly by sampling from those distributions during forward passes is adopted, thereby obtaining concrete weights and biases and adjusting the parameters of the distributions via back-propagation. Deep Gaussian processes are chosen as a deep model based on the widely popular Gaussian processes for probability estimation. Functional neural processes, via a neural architecture, aim to model what Gaussian processes perform adeptly. In our case, we favoured the functional neural processes over other types of neural processes as they are specifically designed to work with a single dataset, while conditional neural processes and (attentive) neural processes truly excel with a large number of diverse datasets for pre-training and a small dataset for training. We have previously excluded the methods that only provide an estimate of the aleatoric uncertainty and opted for methods offering epistemic uncertainty estimates; however, this does not imply that aleatoric uncertainty should be disregarded. It is also a part of the general confidence level of a prediction. If there is uncertainty in the training data, it should be presumed that the predicted situation is uncertain to the same degree. The concept is akin to that used for mixture density networks, but we assume the distribution of the data to be representable as a single Gaussian. To instruct our models to learn the uncertainty in the data, we require them to predict two values: the mean and the standard deviation of a Gaussian distribution. To train the models, we maximize the value of the probability density function at the training data points (i.e., the probability of the training data being in the output distribution). For most methods that provide an estimate of the epistemic uncertainty, a number of samples are drawn, and the parameters of the distribution of the epistemic uncertainty are inferred from the distribution of the samples.

IV. Experimental Analysis

The training and testing of our models utilize energy load data for Germany, obtained from Open Power Systems Data[1]. Specifically, we employ the load data from the ENTSO-E Power Transparency platform (refer to Fig. 1), which has a resolution of 15 minutes. The dataset comprises approximately 140,000 load values, with a fifth of this data (roughly equivalent to one consecutive year) designated as the test set. The remaining data is utilized for hyper-parameter selection via cross-validation and training of the final models. This robust dataset facilitates the development and evaluation of

[1] https://open-power-system-data.org/

our probabilistic deep learning models for energy load forecasting.

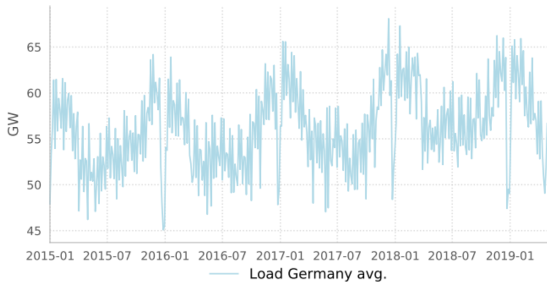

Fig. 1. Information used in scenarios for load forecasting. Germany's load data from early 2015 to early 2019. For clarity, resampled at 4-day intervals

A. *Load Forecasting Analysis*

In our evaluations, we utilized a test set comprising approximately 30,000 data points that were not part of the training data. Additionally, we conducted an assessment of the epistemic uncertainty capabilities of the methods, particularly in relation to anomalous data. It's important to note that while we utilized scaling and de-seasonalization during pre-processing, all values and plots were generated following an inverse transformation of the predicted values back to their original scale.

1) *Short-Term Forecasting*

The short-term forecasting scenario is evaluated first, wherein the models are trained to predict the subsequent value in the electricity load time series. The Probability Integral Transform (PIT) histograms, illustrated in Fig. 2a, reveal that the reference quantile regression model closely approaches a uniform distribution, indicating good probabilistic calibration. However, all models exhibit overdispersion, characterized by bell-shaped PIT histograms, suggesting excessive variability in predictions. The PIT histograms also reveal biases in some models, such as the simple neural network, functional neural process, and deep Gaussian process, which tend to over- or underestimate. Upon recalibration, the PIT histograms (refer to Fig. 2b) exhibit properties akin to a uniform PIT, indicating improved probabilistic calibration. The recalibrated models, with the exception of the simple reference neural network, which retains a slight bias, demonstrate satisfactory calibration. All subsequent results are based on the recalibrated models, ensuring a more accurate assessment of their performance.

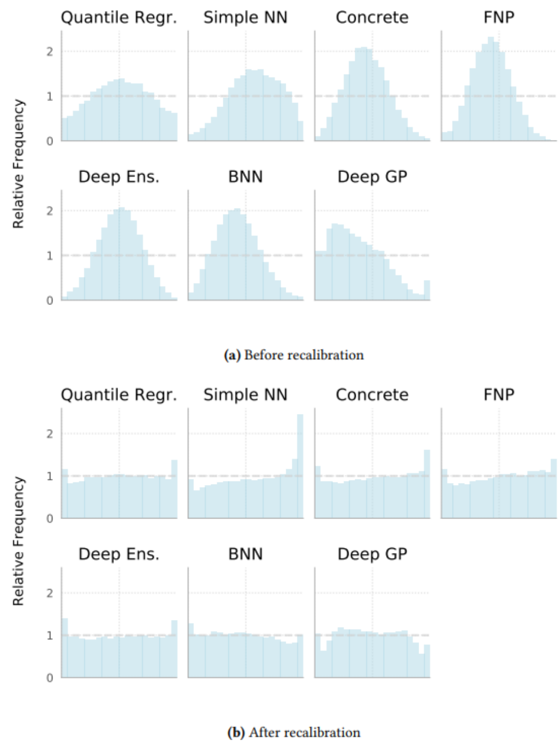

Fig. 2. PIT histograms showing the probability of the various approaches tested in the short-term load forecasting scenario. Displaying the calibration using probability both before and after recalibration. A well-calibrated forecaster's ideal uniform PIT histogram is represented by the grey line in each plot

The box plots in Fig. 3 illustrate the sharpness of the predictions, representing the models' general confidence. The plots display the 90% confidence interval widths of the methods, with optimal sharpness achieved when the model is both confident and accurately representative of the data's real distribution. Deep ensembles and concrete dropouts exhibit the sharpest predictions, with deep ensembles being generally sharper but with a skewed distribution. Similar behaviour is observed for Bayesian neural network and deep Gaussian process models. The metrics in Table I, including Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Continuous Ranked Probability Score (CRPS), confirm these observations, with lower values indicating better performance. Deep ensembles perform the best, followed by concrete dropout and quantile regression. Additionally, Table II compares the training and prediction times for each method, with main models and simple reference neural network requiring between 2h and 5h to train and relatively short prediction times, except for functional neural processes.

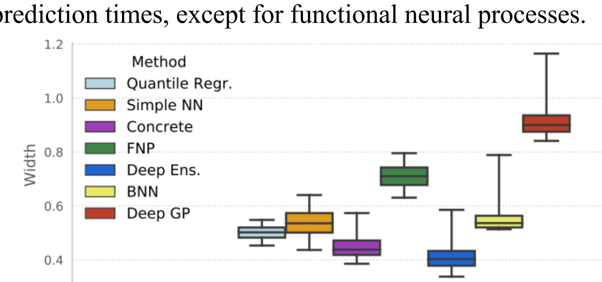

Fig. 3. Distribution of the model predictions' 90% confidence interval widths for the short-term load forecasting scenario following recalibration

TABLE I

FOR THE SHORT-TERM LOAD FORECASTING SCENARIO, RMSE, MAPE, AND CRPS

Model	RMSE	MAPE [%]	CRPS
Deep Ensembles	0.398	0.409	0.170
Concrete	0.413	0.443	0.183
Quantile Regression	0.464	0.462	0.193
BNN	0.471	0.494	0.206
Simple NN	0.496	0.549	0.232
FNP	1.083	0.839	0.331
Deep GP	1.654	1.276	0.494

TABLE II
THE AMOUNT OF TIME NEEDED TO FORECAST AND TRAIN THE SHORT-TERM LOAD FORECASTING SCENARIO MODELS

Model	Train Time [min]	Predict Time [s]
Quantile Regression	0.366	$6.542 \cdot 10^{-2}$
FNP	130.103	161.449
Deep GP	171.437	0.881
Deep Ensembles	198.978	1.748
Simple NN	213.891	0.762
BNN	269.421	9.820
Concrete	287.459	13.580

2) *Day-Ahead Forecasting*

The day-ahead forecast sting scenario is evaluated, where models predict a 24-hour-ahead electricity load without short-term lagged variables. PIT histograms before and after recalibration are presented in Figs. 4a and 5.5b, respectively. Despite recalibration, probabilistic calibration results are inferior to the short-term scenario, with only quantile regression, Bayesian neural network, and deep Gaussian process achieving satisfactory calibration. Most models exhibit a bias to the right and under-dispersion, indicating overconfidence in predictions. Recalibration is less effective compared to the short-term scenario, but initial calibration was already relatively good. Marginal calibration plots (refer to Fig. 5) show similar results for most models, except functional neural process, which performs worse. Accuracy metrics (refer to Table III) rank Bayesian neural networks as the best, closely followed by concrete dropout, reference models, and deep ensembles. The deep Gaussian process performs relatively well, unlike in the short-term case. However, the deep ensemble would rank lower if evaluated by MAPE or RMSE. Sharpness (refer to Fig. 6) is highest for deep ensemble, followed by simple neural network, Bayesian neural network, and concrete dropout. Computation times (refer to Table IV) are similar to the short-term case, with some models requiring less training time. The marginal calibration plots in Fig. 5 exhibit similar patterns for quantile regression, simple neural network, concrete dropout, deep ensemble, Bayesian neural network, and deep Gaussian process model, while functional neural process performs notably worse. In terms of accuracy metrics for day-ahead forecasting, the Bayesian neural network excels, closely followed by concrete dropout, reference models, and deep ensembles, with the deep Gaussian process also performing well. However, the deep ensemble's ranking varies across metrics, surpassing quantile regression in MAPE and RMSE but lagging in probabilistic calibration. The deep ensemble achieves the sharpest results (refer to Fig. 6), followed closely by a simple neural network, Bayesian neural network, and concrete dropout, while the deep Gaussian process produces less sharp results. Computation times (refer to Table IV) are similar to the short-term case, with a simple neural network, deep ensemble, and concrete dropout requiring significantly less training time.

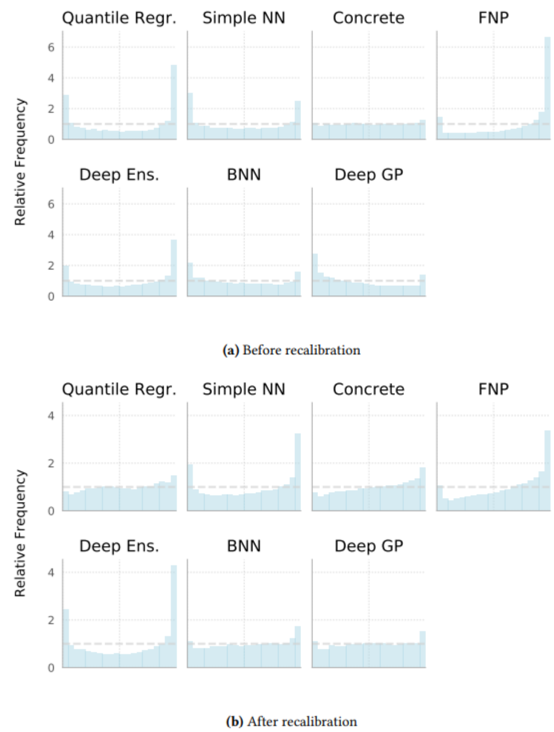

Fig. 4. PIT histograms showing the probability of the various approaches examined in the day-ahead load forecasting scenario. Displaying the calibration using probability both before and after recalibration. For a forecaster with proper calibration, an ideal uniform PIT histogram is shown by the grey line in each plot

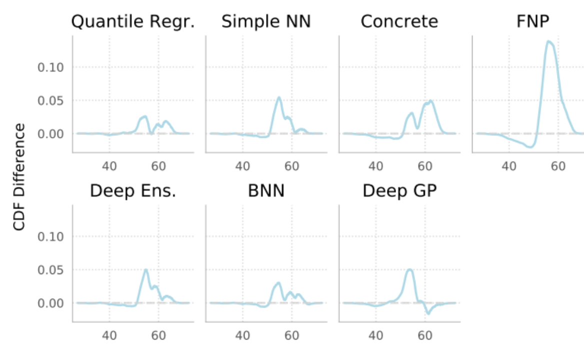

Fig. 5. Plots that assess the techniques' marginal calibration for the day-ahead load forecasting scenario following recalibration. The graphic displays the difference between the empirical distribution of the data' CDF and the average prediction CDF

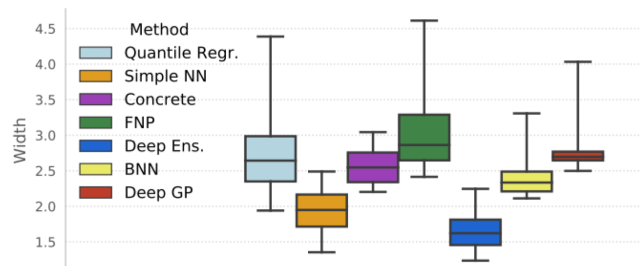

Fig. 6. Distribution of the model predictions' 90% confidence interval widths for the day-ahead load forecasting scenario following recalibration

TABLE III
FOR THE DAY-AHEAD LOAD FORECASTING SCENARIO, RMSE, MAPE, AND CRPS

Model	RMSE	MAPE [%]	CRPS
BNN	1.838	2.362	0.969
Concrete	1.905	2.475	1.010
Simple NN	1.813	2.427	1.026
Quantile Regression	2.108	2.619	1.054
Deep Ensembles	1.866	2.478	1.063
Deep GP	2.536	2.958	1.203
FNP	3.147	3.872	1.624

TABLE IV
THE AMOUNT OF TIME NEEDED TO FORECAST AND TRAIN THE DAY-AHEAD LOAD FORECASTING SCENARIO MODELS

Model	Train Time [min]	Predict Time [s]
Quantile Regression	0.284	$6.170 \cdot 10^{-2}$
Concrete	18.910	12.624
Deep Ensembles	59.091	1.863
Simple NN	64.535	0.478
FNP	131.523	160.869
Deep GP	173.722	0.375
BNN	257.258	10.208

V. Conclusion and Future Works

This study presents a comprehensive examination of probabilistic deep learning methods for energy time series forecasting, evaluating various approaches to acquire probabilistic predictions. The research reveals that most models exhibit similar capabilities, with some demonstrating superior performance. Notably, the best models surpass reference models like quantile regression and simple neural networks, particularly in probabilistic forecasting. However, certain models, such as Gaussian processes and deep Gaussian processes, struggle with scalability. The study highlights the significance of probabilistic predictions in interpreting and evaluating model predictions and identifies areas for future research, including enhancing functional neural processes and deep Gaussian processes. Furthermore, the study suggests exploring alternative variants of neural processes, recurrent methods, and attention-based models to advance the field. Overall, this study provides a thorough overview of probabilistic deep learning methods for energy time series forecasting and identifies opportunities for further research.

VI. Declarations

A. ***Funding:*** No funds, grants, or other support was received.

B. ***Conflict of Interest:*** The authors declare that they have no known competing for financial interests or personal relationships that could have appeared to influence the work reported in this paper.

C. ***Data Availability:*** Data will be made on reasonable request.

D. Code Availability: Code will be made on reasonable request.

References

[1] S. Wazir, G. S. Kashyap, and P. Saxena, "MLOps: A Review," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.10908v1

[2] G. S. Kashyap *et al.*, "Revolutionizing Agriculture: A Comprehensive Review of Artificial Intelligence Techniques in Farming," Feb. 2024, doi: 10.21203/RS.3.RS-3984385/V1.

[3] N. Marwah, V. K. Singh, G. S. Kashyap, and S. Wazir, "An analysis of the robustness of UAV agriculture field coverage using multi-agent reinforcement learning," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 2317–2327, May 2023, doi: 10.1007/s41870-023-01264-0.

[4] S. Wazir, G. S. Kashyap, K. Malik, and A. E. I. Brownlee, "Predicting the Infection Level of COVID-19 Virus Using Normal Distribution-Based Approximation Model and PSO," Springer, Cham, 2023, pp. 75–91. doi: 10.1007/978-3-031-33183-1_5.

[5] M. Kanojia, P. Kamani, G. S. Kashyap, S. Naz, S. Wazir, and A. Chauhan, "Alternative Agriculture Land-Use Transformation Pathways by Partial-Equilibrium Agricultural Sector Model: A Mathematical Approach," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.11632v1

[6] H. Habib, G. S. Kashyap, N. Tabassum, and T. Nafis, "Stock Price Prediction Using Artificial Intelligence Based on LSTM– Deep Learning Model," in *Artificial Intelligence & Blockchain in Cyber Physical Systems: Technologies & Applications*, CRC Press, 2023, pp. 93–99. doi: 10.1201/9781003190301-6.

[7] S. Naz and G. S. Kashyap, "Enhancing the predictive capability of a mathematical model for pseudomonas aeruginosa through artificial neural networks," *Int. J. Inf. Technol. 2024*, pp. 1–10, Feb. 2024, doi: 10.1007/S41870-023-01721-W.

[8] Y. Wang, S. M. Resnick, and C. Davatzikos, "Spatio-temporal analysis of brain MRI images using hidden Markov models," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Med Image Comput Comput Assist Interv, 2010, pp. 160–168. doi: 10.1007/978-3-642-15745-5_20.

[9] T. U. Rehman, M. S. Mahmud, Y. K. Chang, J. Jin, and J. Shin, "Current and future applications of statistical machine learning algorithms for agricultural machine vision systems," Jan. 01, 2019, *Elsevier*. doi: 10.1016/j.compag.2018.12.006.

[10] A. Indika, N. Warusamana, E. Welikala, and S. Deegalla, "Ensemble Stock Market Prediction using SVM, LSTM, and Linear Regression," *TechRxiv*, pp. 1–6, Oct. 2021, doi: 10.36227/TECHRXIV.16626019.V1.

[11] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," *IEEE Access*, vol. 10, pp. 39700–39715, 2022, doi: 10.1109/ACCESS.2022.3166891.

[12] F. Galbusera, G. Casaroli, and T. Bassani, "Artificial intelligence and machine learning in spine research," Mar. 01, 2019, *John Wiley & Sons, Ltd*. doi: 10.1002/jsp2.1044.

[13] P. Zhang, M. Wu, H. DInkel, and K. Yu, "DEPA: Self-Supervised Audio Embedding for Depression Detection," in *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2021, pp. 135–143. doi: 10.1145/3474085.3479236.

[14] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, Cham, 2016, pp. 451–466. doi: 10.1007/978-3-319-46478-7_28.

[15] M. Bicego, A. Farinelli, E. Grosso, D. Paolini, and S. D. Ramchurn, "On the distinctiveness of the electricity load profile," *Pattern Recognit.*, vol. 74, pp. 317–325, Feb. 2018, doi: 10.1016/j.patcog.2017.09.039.

[16] B. Ibragimov, R. Korez, B. Likar, F. Pernus, L. Xing, and T. Vrtovec, "Segmentation of Pathological Structures by Landmark-Assisted Deformable Models," *IEEE Trans. Med. Imaging*, vol. 36, no. 7, pp. 1457–1469, Jul. 2017, doi: 10.1109/TMI.2017.2667578.

[17] M. Nasri, A. Helali, H. Sghaier, and H. Maaref, "Adaptive image compression technique for wireless sensor networks," *Comput. Electr. Eng.*, vol. 37, no. 5, pp. 798–810, Sep. 2011, doi: 10.1016/j.compeleceng.2011.08.001.

[18] S. E. Tranter, K. Yu, G. Evermann, and P. C. Woodland, "Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2004. doi: 10.1109/icassp.2004.1326095.

[19] M. Dutta *et al.*, "Monitoring Root and Shoot Characteristics for the

Sustainable Growth of Barley Using an IoT-Enabled Hydroponic System and AquaCrop Simulator," *Sustain.*, vol. 15, no. 5, p. 4396, Mar. 2023, doi: 10.3390/su15054396.

[20] G. S. Kashyap, A. Siddiqui, R. Siddiqui, K. Malik, S. Wazir, and A. E. I. Brownlee, "Prediction of Suicidal Risk Using Machine Learning Models," Dec. 25, 2021. Accessed: Feb. 04, 2024. [Online]. Available: https://papers.ssrn.com/abstract=4709789

[21] P. Kaur, G. S. Kashyap, A. Kumar, M. T. Nafis, S. Kumar, and V. Shokeen, "From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility," Feb. 2024, Accessed: Mar. 21, 2024. [Online]. Available: https://arxiv.org/abs/2402.16142v1

[22] F. Alharbi and G. S. Kashyap, "Empowering Network Security through Advanced Analysis of Malware Samples: Leveraging System Metrics and Network Log Data for Informed Decision-Making," *Int. J. Networked Distrib. Comput.*, pp. 1–15, Jun. 2024, doi: 10.1007/s44227-024-00032-1.

[23] G. S. Kashyap *et al.*, "Detection of a facemask in real-time using deep learning methods: Prevention of Covid 19," Jan. 2024, Accessed: Feb. 04, 2024. [Online]. Available: https://arxiv.org/abs/2401.15675v1

[24] G. S. Kashyap, K. Malik, S. Wazir, and R. Khan, "Using Machine Learning to Quantify the Multimedia Risk Due to Fuzzing," *Multimed. Tools Appl.*, vol. 81, no. 25, pp. 36685–36698, Oct. 2022, doi: 10.1007/s11042-021-11558-9.

[25] M. Sharma, P. Bangotra, A. S. Gautam, and S. Gautam, "Sensitivity of normalized difference vegetation index (NDVI) to land surface temperature, soil moisture and precipitation over district Gautam Buddh Nagar, UP, India," *Stoch. Environ. Res. Risk Assess.*, vol. 36, no. 6, 2022, doi: 10.1007/s00477-021-02066-1.