# Creating Non-Trivial Wildfire Prediction Models

By Jared Hayes, Cayden Wilson, Aubay Azzarouk, Andrew Wenzel

## Abstract

With advancements in machine learning and the increasing frequency of wildfires in the state of California, it is natural to wonder if a model can be made to predict when and where these disasters will occur. While some models exist, none are sufficient for real-world use. In this study, we implement an Enhanced U-Net Convolutional Long Short-Term Memory (U-LSTM) neural network architecture for wildfire prediction, trained on multimodal data that was extracted directly from Google Earth Engine. We were able to reproduce similarly high accuracies to the gold standard model we adapted, which raised concerns about the validity of certain aspects of this prediction problem. We propose a new method of accuracy incorporating multiple prediction tasks as well as a model for compiling more robust input measures and ground truths. This approach not only highlights the limitations of current evaluation standards but also lays the groundwork for more meaningful and actionable wildfire prediction systems in the future.

## Introduction

Wildfire prediction models have advanced significantly in recent years, incorporating machine learning techniques to process multi-modal data from satellite imagery, weather stations, and topographical information. These models typically generate spatio-temporal heatmaps that estimate fire probability across regions over time. Traditional classification metrics such as accuracy, precision, recall, and F1 score are commonly used to evaluate these predictions. Additionally, recent research by Bhowmik et al. (2023) utilized a modified mean squared error approach as their accuracy metric for their U-Convolutional Long Short-Term Memory (ULSTM) neural network, which achieved impressive results in predicting California wildfires.

Despite these advancements, a significant gap exists in how we evaluate wildfire prediction models. Current metrics fail to account for several critical operational factors: (1) the asymmetric costs of false negatives versus false positives, (2) the varying importance of fires based on their size, and (3) the complexity of ground truth burn maps. Standard classification metrics treat all errors equally, while operational reality suggests that missing a large fire has far greater consequences than raising a false alarm, and predicting a fire two weeks in advance is more valuable than one day in advance. Without metrics that incorporate these factors, model selection and development may be guided by inappropriate evaluation criteria, potentially reducing their real-world utility.

## Background

When examining the problem of predictive algorithms for wildfires, we came across a model that showed exceedingly high accuracy (97%) when predicting fire intensity, and this model measured intensity with mean absolute error. Initially the goal we had was to replicate this process and try to improve the model to work on a less complex dataset. This did work, with our model achieving similar results to those of the parent paper; however, the process raised warning flags in our heads about how the accuracy was being calculated. If it was really true that multiple models could predict wildfires with near-perfect accuracy, why were these models not being utilized by states like California? This question led us to investigate different approaches to analyzing the accuracy of models making predictions in this context.

## Research Question and Approach

This paper addresses a fundamental question: How can we develop an evaluation metric for wildfire prediction models that better aligns with operational priorities and practical utility? To answer this question, we propose a comprehensive evaluation framework that incorporates asymmetric error penalties, burn area mapping, and a spatial-temporal model architecture.

## Related Work

### Multi-Modal Prediction Systems

As stated previously, there are multiple existing models that attempt to address the issue of wildfire prediction. However, the vast majority of these models use relatively basic methods in order to calculate the accuracy of their model. The primary paper we referenced throughout this project is *A multi-modal wildfire prediction and early-warning system based on a novel machine learning framework* which we refer to as the Stanford paper. This paper served as the baseline for the model we built. However, while the reported accuracy of 97% is remarkable, the authors only used binary cross entropy as their error metric and used improper ground truth data. While this is definitely a good starting point in the context of the problem, we do not believe it fully captures the intricacies of such a complex problem. The paper itself admits that it is really only effective at predicting large scale fires and falls short when predicting smaller ones. The issue is that because these fires are so much smaller in comparison to the large ones, this shortcoming is not reflected in the accuracy statement.

As the paper claims to have a novel 97% accuracy but only in the last paragraph of the results does it discuss how on all the fires it only had an accuracy of around 42%. This appears, to some extent, as a misrepresentation of results because as we showed it is a relatively easy problem to create heat maps at a very high accuracy at a low resolution which is similar to creating heatmaps at a very high accuracy for large scale fires. As well, using modified mean-squared error for accuracy fails to encapsulate the complexity of the problem because fires are a rare occurrence which means that is easy to build a model that can predict the accuracy of the fire not happening, but it is much harder to predict the occurrence of the fires and characteristics of the fires such as spread. Also, modified mean-square error treats all the pixels as though they are independent which is not an entirely true assumption to be made. It is our goal to improve upon this existing work by establishing a method of acquiring accuracy that better displays how effective the model is in order to make these predictions viable for real-world use cases.

## Data Science and Machine Learning Applications

Jain et al. (2020) conducted a comprehensive review of machine learning applications in wildfire science and management, noting diverse approaches being adopted across prediction, detection, and assessment phases [2]. They identified the need for more sophisticated evaluation metrics that address the spatial and temporal aspects of wildfire prediction.

Alfred (2022) compared various data science techniques for wildfire investigations, evaluating Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and decision tree models [3]. His study found that different algorithms performed optimally under different conditions and data scales, suggesting that a single accuracy metric may be insufficient for comprehensive evaluation.

### Remote Sensing and Geospatial Integration

Koubarakis et al. (2013) described a real-time wildfire monitoring service using scientific database and linked data technologies [4]. Their system integrated satellite imagery with geospatial data to detect hotspots and monitor fire spread. While effective, their evaluation focused primarily on detection accuracy rather than comprehensive prediction metrics.

Naganathan et al. (2016) explored wildfire predictions using fused datasets, combining meteorological data with fire databases [5]. Their work highlighted the challenges in evaluating model performance when using integrated data sources, suggesting that binary accuracy metrics alone may not adequately represent prediction quality.

### Regional Ignition Probability Models

Larjavaara et al. (2004) examined climate-influenced variation in forest fire ignition probability across different regions [6]. Their work demonstrated the importance of regional context in prediction models and the need for evaluation metrics that account for geographical variations in fire behavior.

### Ignition Modeling

Reszka et al. (2012) presented a methodology for estimating ignition delay times in forest fire modeling [7]. Their approach focused on specific physical aspects of fire initiation rather than the broader prediction challenge, highlighting the multi-faceted nature of wildfire prediction that binary accuracy metrics might oversimplify.

## Methodology

In our research, we utilized Google Earth Engine (GEE) to systematically collect and process environmental data relevant to wildfire prediction across California counties. Our data pipeline initializes GEE and establishes connections to Google Drive for efficient data storage, then extracts multiple environmental variables between 2013-2024 including vegetation indices (NDVI) from MODIS satellite imagery, temperature and dew point measurements from ERA5 climate datasets, wind components, and critically, Fire Radiative Power (FRP) from VIIRS satellite data which serves as our primary fire intensity metric. The script harmonizes these diverse data sources to consistent resolutions—10km for environmental factors and 375m for fire intensity measurements—and converts all layers to Float32 data type to ensure computational compatibility. Processing occurs in monthly batches with controlled parallel execution (limited to three simultaneous tasks) to respect GEE's operational constraints while maximizing efficiency. This data preparation framework addresses key challenges in wildfire prediction modeling, particularly the integration of multi-modal data with varying temporal and spatial resolutions, and creates standardized, temporally consistent datasets that combine environmental factors known to influence wildfire behavior, providing the foundation for our LSTM-based predictive model.

For this paper, we are using a slightly altered version of the model from [1]. This model combines two main components: a U-Net and an LSTM. The U-Net model is the first step, and it is applied to the satellite imagery to extract spatial features across multiple scales. The U-Net architecture employs an encoder path that progressively downsamples the input while increasing feature depth, followed by a decoder path that up-samples the features while incorporating skip connections from the encoder, allowing the network to retain fine-grained spatial information. The LSTM is the second stage of the model, and it learns the temporal component by processing the sequence of extracted features across time steps, enabling the model to identify patterns in how fires develop and spread over time. After the images have been processed and compiled into time-series format, the ConvLSTM cells capture spatial-temporal dependencies that are crucial for predicting wildfire progression.

Our implementation employs an Enhanced U-Net Convolutional Long Short-Term Memory (ULSTM) neural network architecture for wildfire prediction. The model processes temporal sequences of satellite imagery and environmental data through a robust dataset handler that accommodates multi-channel geospatial TIF files with varying numbers of channels. This dataset handler automatically normalizes input data and calculates vegetation indices like NDVI and NBR when available, providing additional features that have proven valuable for fire prediction. Our implementation incorporates chronological data splitting to properly evaluate temporal prediction performance, avoiding the data leakage that can occur with random splitting, where future fire patterns might inadvertently influence the training of models that should predict these patterns.

The output of the model is a value between 0 and 1 for each tile on the grid. Values closer to 1 indicate a higher likelihood of there being a fire while values closer to 0 indicate a lower likelihood. We then compared these values to a threshold, predicting 1 (there is a fire) for values above the threshold and 0 (there is no fire) for values below it. These binary responses are what will be used to calculate the final accuracy metrics. We utilized Binary Cross Entropy (BCE) as our loss function, which provides a straightforward approach to training the model for binary classification of fire occurrence. Our implementation includes comprehensive evaluation metrics and visualization tools to assess model performance. Given the highly imbalanced nature of wildfire data (where fire pixels are vastly outnumbered by non-fire pixels), we track precision, recall, and F1 scores in addition to standard loss measurements. The metrics were based on the number of correct pixels versus the number of incorrect pixels, with precision measuring the ratio of correctly predicted fire pixels to total predicted fire pixels, recall measuring the ratio of correctly predicted fire pixels to actual fire pixels, and F1 score providing a harmonic mean of these two values. The code generates heat map visualizations comparing predicted fires against ground truth, with color-coding to highlight true positives, false positives, and false negatives, providing intuitive understanding of model performance.

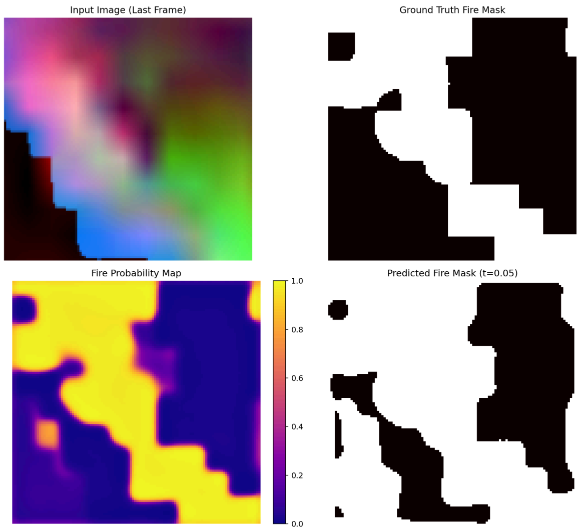

Figure 1. Visualizations Produced by our model

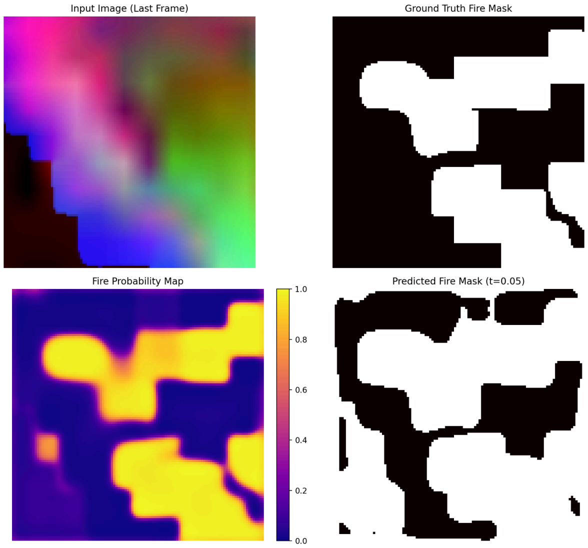

Figure 2. Continued Visualizations from our model

Figure 1 and Figure 2 show some visualizations of our model and the fire probability heatmaps it was producing. In Figure 1 and Figure 2 the top left image shows the last input image which was fed to our model. The bottom left image shows the fire probability heatmap. The top left image shows the ground truth fire mask which was pulled from the MODIS satellite which is showing the fire mask pulled from the satellite. While the bottom right image is what our model predicted as the fire mask from the satellite. F1 score, recall, and precision were measured by comparing the ground truth fire mask to the fire probability map pixel by pixel.

As seen from the images above, too many simplifications were made in attempting to reproduce the Stanford model, such as the data resolution and implementation of ground truths, which limited the difficulty of the problem when compared to the complexity of the model. However, it does show that the inherent problem of creating a probability heatmap can be accomplished to a high degree of accuracy at large scale with a small scale of resolution by a combination of an U-net and a LTSM.

This is most likely because the U-Net architecture excels at identifying spatial patterns and features in the input imagery, while the LSTM component effectively captures the temporal evolution of these features across sequential satellite images. The combination addresses both the spatial and temporal dimensions of wildfire behavior, which are inherently linked. While our simplified implementation achieves strong results at coarser resolutions, it highlights a critical insight for wildfire prediction models: the evaluation metrics need to be carefully considered beyond raw accuracy. As demonstrated in our visualizations, even a model with high overall accuracy may still miss critical fire events or produce false alarms in vulnerable areas, underscoring our paper's central argument that specialized evaluation frameworks incorporating asymmetric error penalties and contextual fire importance are essential for meaningful assessment of wildfire prediction capabilities.

## Limitations in Current Evaluation Methods

The current literature reveals a significant gap in evaluation methodology. While many models report high accuracy values, these metrics often fail to capture:

1. Limited ground truth data
2. Performance differences between large and small fire events
3. Spatial precision of predictions
4. Economic and human impact dimensions of prediction value

The primary issue we encountered was in generating ground truths for the affected wildfire areas. After our initial testing, it became clear that we did not have an accurate way to represent the burn maps that would be needed for model training. Thus, we re-examined the methods used by Bhowmik et al. (2023) and discovered that the strategy employed was insufficient to measure accuracy. Bhowmik et al. generate circles as ground truths of the fire data based on the fire's intensity, but fires do not always spread in a perfect circle.

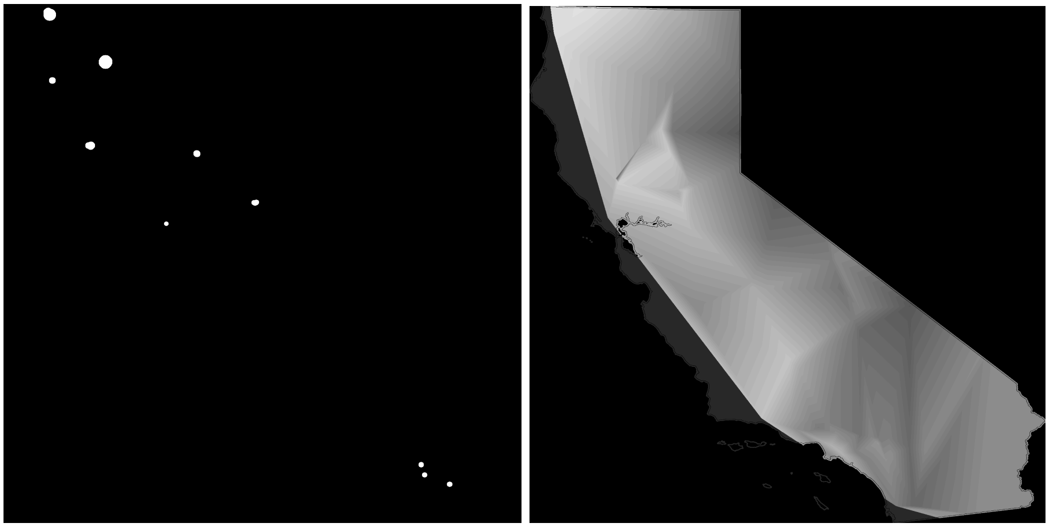

Figure 3. Ground truth (left) and example interpolated feature map (right) used by Bhowmik et al. (2023)

Binary cross-entropy as a loss function, while convenient, may overstate model effectiveness due to the imbalanced nature of wildfire occurrences, where large fires are rarer but have disproportionately large importance in the accuracy calculation.

This gap presents an opportunity to develop more nuanced evaluation frameworks that better represent real-world prediction utility and can guide practical implementation of wildfire prediction systems.

## Measuring Accuracy

Bhowmik et al. (2023) measures the accuracy of predictions for individual tiles with mean absolute error. The metric being evaluated is fire intensity modeled by Fire Radiative Power (FRP). This is the primary accuracy metric of the model in [1], but this interpretation of accuracy faces challenges for a variety of reasons.

When developing the accuracy function for this problem, there were a lot of things that we needed to consider. Contextually, there is not one trivial "correct" way of defining error and accuracy when making these predictions. Obviously it is important not to miss fires, so the model needs to avoid false negatives at all costs. However, overprediction would also be negative as it could result in misallocation of resources or unnecessary panic. In addition, when predicting spread of a large scale or small scale fire, connectivity is important to consider. If the model predicts a fire one tile next to where it actually is and misses that tile, the resulting impact would likely not be severe in terms of actually fighting the fire. Similarly, regionality is an important concept. If resources are directed to a particular region, regardless of if the exact location of the fire was correct, that would still have a positive impact on the fighting of the fire and on any potential evacuation response. Finally, it is important to consider the significance of the individual tiles in terms of potential damages and threat to populations. Missing a tile in the middle of a mountain range where nobody lives would likely not cause harm, while missing a tile adjacent to a heavily populated area or area of significant infrastructure would be much more harmful.

### Fire Intensity

Fire Radiative Power (FRP) is a variable defined by the instantaneous output of energy from a fire, and it has been found to be highly correlated with biomass loss and smoke output during wildfire events. Several papers utilize this metric as the The problem with FRP is that measurements can often be misleading or missing entirely during large wildfire events despite being an indicator of intensity [8]. Therefore, it is not sufficient to compute the model's accuracy using this metric alone.

### Fire Spread Time

Including measurement of the speed of fire spread is critical to allocating firefighting resources and response teams. Jiang et al. (2022) leverage a graph model to predict both the perimeter and velocity of the fire spread, but they do not specify an overall measure that incorporates intensity, spread perimeter, and speed. Furthermore, there is limited research aimed towards predicting factors like the speed of the fire, so it is difficult to make comparisons between models. We include fire spread time to be measured within the accuracy function so responders can adjust their plans accordingly.

### Composite Accuracy Measure

A simplified way of interpreting the overall accuracy quantification can be described as aggregating the accuracies of each individual predictive task. This overarching accuracy measure captures the model's reliability in predicting not only the FRP (i.e. fire intensity), but other response-relevant factors like perimeter, speed, and risk to civilian infrastructure. Those factors can then be weighted relevant to their importance to wildfire response teams, and any number of factors can be used depending on the specific use case. A trivial example of such a function can be found below.

$$Accuracy_{Total} = w_1\, Accuracy_{Intensity} + w_2\, Accuracy_{Speed} + \ldots$$

This flexible formulation enables stakeholders to tailor the accuracy metric to prioritize the most mission-critical predictions for effective wildfire management up to two weeks in advance.

## Future Works

While we believe that this a good step towards improving implementation of wildfire prediction models into real-world use cases, there is still plenty of room for future research. A critical area advancement is the refinement of accuracy metrics to encompass the full nature of wildfire behaviors. Further complexities could still be incorporated into defining the accuracy, and more development on models is necessary to meet the standards of these accuracy functions.

Moreover, even when these multi-modal predictive models demonstrate high accuracy, their deployment by state agencies like those in California can be hindered by factors such as data variability, environmental unpredictability, and computational constraints, underscoring the need for continued research to ensure that predictive tools are both fully interpretable and actionable for key decision-makers.

A solution is needed that employs the use of an algorithm like the one presented by Ban et al. (2020). Once enough extracted burn maps are compiled for California, then an effective training process can occur with the burn maps acting as ground truth data.

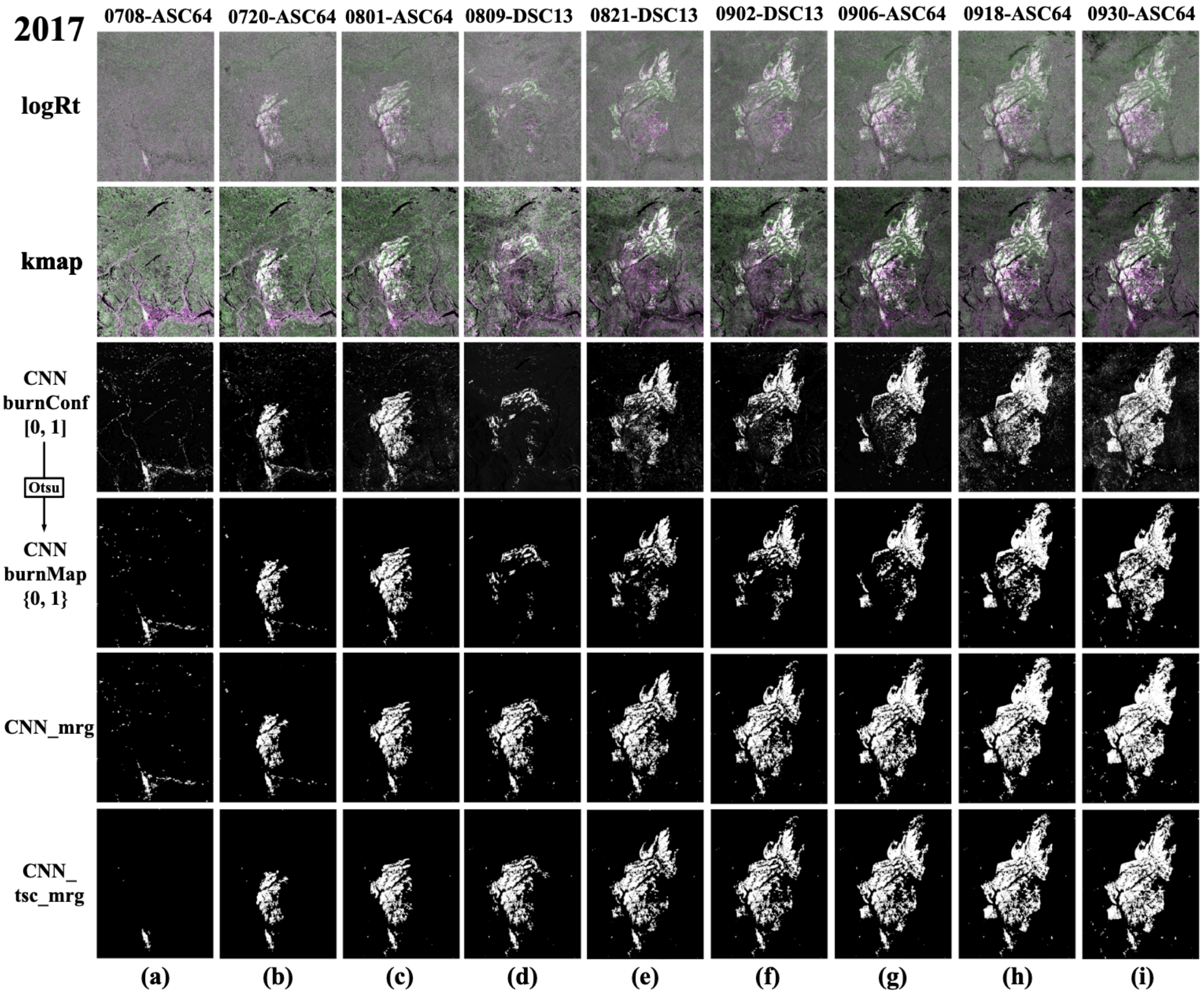

Figure 4. Various methods for burn map extraction employed by Ban et al. (2020)

## Discussion/Conclusion

We explored a U-Net LSTM architecture for wildfire prediction and revealed shortcomings of the current literature. One proposed method for generating ground truth data is through algorithms that extract accurate burn maps. Additionally, a multi-task prediction model enables emergency responders, civilians, and resource planners to evaluate models based on their own needs. Our work highlights the need for more robust wildfire datasets to enable multi-tasked prediction models.

## References/Bibliography

[1] Bhowmik, R. T., Jung, Y. S., Aguilera, J. A., Prunicki, M., & Nadeau, K. (2023). A multi-modal wildfire prediction and early-warning system based on a novel machine learning framework. Journal of Environmental Management.

[2] Jain, P., Coogan, S.C.P., Subramanian, S.G., Crowley, M., Taylor, S., & Flannigan, M.D. (2020). A review of machine learning applications in wildfire science and management. Environmental Reviews, 28(4), 478-505.

[3] Alfred, R. (2022). Comparative Study - The Application of Data Science to Wildfire Investigations. Technoarete Transactions on Advances in Data Science and Analytics, 1(2), 13-20.

[4] Koubarakis, M., Kontoes, C., & Manegold, S. (2013). Real-Time Wildfire Monitoring Using Scientific Database and Linked Data Technologies. EDBT/ICDT, 649-660.

[5] Naganathan, H., Seshasayee, S.P., Kim, J., Chong, W.K., & Chou, J.S. (2016). Wildfire Predictions: Determining Reliable Models using Fused Dataset. Global Journal of Computer Science and Technology, 16(4), 35-46.

[6] Larjavaara, M., Kuuluvainen, T., Tanskanen, H., & Venäläinen, A. (2004). Variation in Forest Fire Ignition Probability in Finland. Silva Fennica, 38(3), 253-266.

[7] Reszka, P., Borowiec, P., Steinhaus, T., & Torero, J.L. (2012). A methodology for the estimation of ignition delay times in forest fire modelling. Combustion and Flame, 159(12), 3652-3657.

[8] Saide, P. E., Krishna, M., Ye, X., Thapa, L. H., Turney, F., Howes, C., & Schmidt, C. C. (2023). Estimating fire radiative power using weather radar products for wildfires. *Geophysical Research Letters*, *50*(21). https://doi.org/10.1029/2023gl104824

[9] Jiang, W., Wang, F., Su, G., Li, X., Wang, G., Zheng, X., Wang, T., & Meng, Q. (2022). Modeling Wildfire Spread with an Irregular Graph Network. *Fire*, *5*(6), 185. https://doi.org/10.3390/fire5060185

[10] Ban, Y., Zhang, P., Nascetti, A. et al. Near Real-Time Wildfire Progression Monitoring with Sentinel-1 SAR Time Series and Deep Learning. Sci Rep 10, 1322 (2020). https://doi.org/10.1038/s41598-019-56967-x