applied sciences

MDPI

Article

Identifying Earthquakes in Low-Cost Sensor Signals Contaminated with Vehicular Noise

Leonidas Agathos, Andreas Avgoustis, Nikolaos Avgoustis, Ioannis Vlachos, Ioannis Karydis and Markos Avlonitis *

Department of Informatics, Ionian University, 49132 Kerkyra, Greece; inf.bdn2203@ionio.gr (L.A.); inf.bdn2202@ionio.gr (A.A.); avgoustis@ionio.gr (N.A.); gvlachos@ionio.gr (I.V.); karydis@ionio.gr (I.K.)
* Correspondence: avlon@ionio.gr; Tel.: +30-266187766

Abstract: The importance of monitoring earthquakes for disaster management, public safety, and scientific research can hardly be overstated. The emergence of low-cost seismic sensors offers potential for widespread deployment due to their affordability. Nevertheless, vehicular noise in low-cost seismic sensors presents as a significant challenge in urban environments where such sensors are often deployed. In order to address these challenges, this work proposes the use of an amalgamated deep neural network constituent of a DNN trained on earthquake signals from professional sensory equipment as well as a DNN trained on vehicular signals from low-cost sensors for the purpose of earthquake identification in signals from low-cost sensors contaminated with vehicular noise. To this end, we present low-cost seismic sensory equipment and three discrete datasets that—when the proposed methodology is applied—are shown to significantly outperform a generic stochastic differential model in terms of effectiveness and efficiency.

Keywords: low-cost sensors; deep neural networks; vehicular noise; earthquake measurement; earthquake signal contamination; seismometer

Citation: Agathos, L.; Avgoustis, A.; Avgoustis, N.; Vlachos, I.; Karydis, I.; Avlonitis, M. Identifying Earthquakes in Low-Cost Sensor Signals Contaminated with Vehicular Noise. *Appl. Sci.* **2023**, *13*, 10884. https://doi.org/10.3390/app131910884

Academic Editors: Shiyong Zhou and Ke Jia

Received: 13 September 2023
Revised: 25 September 2023
Accepted: 27 September 2023
Published: 30 September 2023

1. Introduction

Earthquakes are sudden movements along fault lines that release stored elastic energy in rocks, generating seismic waves that propagate throughout the Earth [1]. Seismology is a field abundant with data and is heavily reliant on data analysis. Each day witnesses numerous earthquakes worldwide with magnitudes exceeding 2.5, which can be felt locally. Additionally, every few days, earthquakes capable of causing structural damage occur [2]. Moreover, there is a continuous occurrence of numerous smaller earthquakes, typically with magnitudes below 2.5, which are too weak to be perceptible but are consistently recorded by modern instruments [3]. These minor seismic events offer valuable insights into the mechanisms of earthquakes [4].

Monitoring earthquakes is important for disaster management, public safety, and scientific research. It enables determining preparatory activities such as timely warnings, evacuation plans, and response strategies in order to mitigate the impact of seismic events, in addition to studying earthquake patterns to accrue valuable insights into Earth's dynamics. To this end, sensor networks, comprising mostly of seismometers and accelerometers placed throughout the globe, play a pivotal role in this effort by continuously collecting seismic data. These networks facilitate real-time monitoring, data analysis, and the development of earthquake prediction models, ultimately enhancing the ability to protect lives and infrastructure, and advance the scientific understanding of earthquake behavior.

The emergence of low-cost sensors represents a significant advancement in environmental monitoring [5], in general, and earthquake monitoring [6], specifically, as their affordability allows for widespread deployment. These sensors offer a plethora of advantages, including increased spatial coverage and dense monitoring networks that allow for

a more comprehensive understanding of environmental conditions. Moreover, their cost-effectiveness allows for easier replacement in the event of malfunction or damage, ensuring continuous data collection without substantial financial burdens. This democratization of sensor technology not only enhances our ability to gather data across vast geographical areas but also empowers communities, researchers, and organizations to address critical environmental and societal challenges with greater precision and efficiency.

Vehicular noise contamination poses a significant challenge in urban environments, where the deployment of low-cost sensors is common. This challenge stems from the ubiquitous presence of vehicles on roads and highways, generating a continuous stream of noise that can impact various aspects of urban life [7]. The deployment of low-cost sensors, while advantageous for monitoring purposes, can also exacerbate the problem by providing a platform for capturing and transmitting this noise. As urban areas continue to grow and traffic congestion increases, the issue of vehicular noise becomes more pronounced, affecting the well-being of residents, wildlife habitats, and overall quality of life [8]. The task of effectively using sensor data to monitor and analyze while accounting for noise contamination presents a complex problem that requires innovative solutions and advanced earthquake signal processing and deep learning techniques [9].

Earthquake identification is of paramount importance due to its significant impact on public safety, emergency response efforts, and disaster preparedness. Earthquakes are natural disasters that can cause widespread destruction, loss of life, and disruption to communities. The accurate and timely identification of earthquakes allows a variety of critical actions that can mitigate their effects and save lives [10]. It is thus fundamental to effective disaster management and empowers individuals, communities, and authorities with the information needed to make informed decisions, take swift actions, and allocate resources efficiently. The consequences of false positives, and negatives, underscore the critical nature of reliable seismic monitoring systems in ensuring public safety and disaster preparedness [11].

Motivation and Contribution

The aforementioned importance of accurate earthquake identification can hardly be overstated, given the effect of earthquakes on so many aspects of life. Moreover, the emergence of low-cost sensory equipment for such identification, and its widespread adoption nowadays, calls for research on the effectiveness and efficiency of this use. This is further exacerbated by the high density of such sensors that frequently are adjacent to publicly accessible road infrastructure which in turn contaminates the signals received by the sensors with vehicular noise. It is, thus, the lack of comprehensive studies addressing the impact of vehicular noise on earthquake signals captured by low-cost sensors that this work aims to address.

In order to address these challenges, this work proposes the use of an amalgamated deep neural network (DNN) composed of (i) a DNN trained on earthquake signals from professional sensory equipment, as well as (ii) a DNN trained on vehicular signals from low-cost sensors, for the purpose of earthquake identification in signals from low-cost sensors contaminated with vehicular noise. The key contributions of this work can be summarised as follows:

- Creation and dissemination of a dataset of vehicular noise measured with low-cost seismic sensor;
- Collection and dissemination of ground truth data from professional seismic measurement equipment;
- Creation of DNNs for the aforementioned dataset approaches and experimentation on their performance;
- Creation and dissemination of a two-fold synchronized dataset: seismic data from a low-cost seismic sensor, and seismic data from a professional seismic sensor. Both sensors are in very close proximity; and

- Amalgamation of the aforementioned DNNs for the identification of earthquakes in signals from low-cost sensors contaminated with vehicular noise and experimentation on the DNN.

The rest of this paper is organized as follows: Section 2 discusses the key recent relevant studies about seismology, seismic waves, sensors for measuring and removing noise from seismic waves, and deep learning for earthquake and vehicle classification. Section 3 presents the proposed methodology and the deep learning classification algorithm utilized in this work. Section 4 details the pre-processing techniques applied to the datasets and the experiments conducted, wherein their respective results are presented and discussed. Finally, this paper is concluded in Section 5.

2. Background and Related Work

In seismology, the foundation of knowledge lies in data analysis, with significant breakthroughs often stemming from the examination of fresh datasets or the creation of novel data analysis techniques [12]. Seismology focuses on the study of earthquakes and associated phenomena, primarily applying the principles of continuous medium mechanics, specifically the theory of elasticity. In contrast, seismic engineering is an applied science concerned with understanding how earthquake-induced motion impacts man-made structures, including buildings and other constructions [13].

Earthquake impacts on both natural and human-made structures are primarily driven by the transfer of energy through seismic waves originating from the earthquake's source. Seismic waves propagate through the Earth and are detected at distant locations using sensitive seismographs. Interpreting seismic records requires an understanding of how seismic waves are generated and propagated, and how recording processes affect them. Advances in seismic instrumentation now allow for accurate digital representation of particle motion across a wide frequency range. However, this necessitates careful consideration of seismic noise, the background irregular ground motion caused by various factors, including human activities and natural phenomena. Occasionally, this background noise is interrupted by organized energy patterns generated by seismic waves from natural or artificial sources. These wave-trains, characterized by distinct arrivals associated with specific propagation paths, become more pronounced with increasing distance from the source. Following the initial body-wave phases, like P (compressional waves) and S (shear waves), there is an increase in record amplitude as surface-guided waves arrive [14].

Over the centuries, seismology has evolved significantly. From Zhang Heng's ancient seismograph in 132 AD to the late 1800s when scientific research on seismology began, progress was slow. It was not until around 1900 that precise measuring instruments, like geophones and seismometers, emerged. These early devices were large, costly, and had limited sensitivity. However, recent advances in micro-electro-mechanical system (MEMS) technology have drastically reduced size and cost while improving sensitivity, making MEMS-based seismic sensors highly promising for their ability to provide reliable measurements across a wide bandwidth [15]. Modern seismographs produce digitized information at varying regular time intervals sent to be analyzed on computers. Many concepts of time series analysis, including filtering and spectral methods, are valuable in seismic analysis [16].

The identification of noise (seismic included) depends on a plethora of parameters and usually requires data analysis while depending on the domain or application, a part of the information may be be treated as noise or useful signal [17]. Seismic noise monitoring systems have been proposed [18] that address continuous traffic noise utilizing raw noise records as well as shear-wave velocity profiles. Prior to seismic wave measurement and identification, noise removal is another important factor that has been addressed [19]. Periodic noise poses a well-documented challenge in the context of seismic wave removal, often originating from sources such as power lines, pump jacks, engine operations, or other forms of interference. It introduces contamination to seismic data and has a notable impact on subsequent data processing and interpretation. The proposed denoising approach

hinges on the sparse representation of periodic noise, enabling its estimation without being influenced by seismic reflections. Consequently, this method effectively reduces periodic noise without compromising the integrity of seismic events. Similarly, the utilization of machine learning algorithms for eliminating random noise in seismic data has emerged as a crucial aspect of seismic analysis [20]. In this work, the authors emphasize that the elimination of random noise from seismic data significantly affects the accuracy of subsequent data processing. They achieve an enhancement in the signal-to-noise ratio of seismic data through the application of a convolutional neural network trained on noise. This not only results in a higher signal-to-noise ratio but also preserves more valuable information.

Given the previously discussed importance of earthquake identification, and thus, early warning systems, the prohibitive cost of high-end ground motion sensors often leaves earthquake-prone areas unable to implement such systems for measuring seismic waves. Low-cost MEMS-based ground motion sensors present a promising solution for creating affordable, yet reliable and sturdy, seismometers. Traditional high-end monitoring systems are highly dedicated measuring systems with high to very high precision of measurement, usually significantly above the monitoring scenario's requirements. The low-cost approach in such monitoring systems attempts to minimize the cost (usually at the level of one to, hopefully, two orders of magnitude) while preserving the precision of measurements within acceptable [21]. The lower cost allows for a higher number of deployed systems and a lower cost per system unit replacement, leading to a denser network of interconnected systems compared to high-end solutions, offering redundancy, expansive spatial measurements, and—utilizing AI methods—the capacity for collective extraction of information. This collective approach yields insights unattainable by unique systems, achieving significantly higher levels of precision compared to unique low-cost systems and rivaling those of non-low-cost systems. The advancements in utilizing low-cost sensors for detecting earthquakes and issuing warnings have shown remarkable progress in recent years. This progress is evident in the expansion of station coverage, the enhancement of data quality, and the broadening scope of applications related to earthquake detection [22]. Real-time seismic signal waves are available to be plotted using ShakeMaps, helping to assess the damage patterns and directivity of rupture. These ShakeMaps plots have proven [23] helpful in establishing the peak ground velocity indicator of damage, and the peak ground acceleration.

Similar to our proposed work for earthquake identification, some research efforts have also been made for the event detection of earthquakes with machine learning algorithms, applying time wave series data analogous to those used for different vehicle types. The implementation of different machine learning algorithms determines the class of automobiles [24] for distinguishing between earthquake and non-earthquake, vandalism vibrations [25], even for event detection, phase identification, and the onset picking time [26]. In all such cases, the results indicate that the use of deep neural networks was superior in distinguishing and provided high classification accuracy during training, as well as in the event and phase detection of earthquakes.

3. Proposed Methodology

Our work proposes the use of an amalgamated deep neural network constituent of a deep neural network trained on earthquake signals from professional sensory equipment, as well as a deep neural network trained on vehicular signals from low-cost sensors. These sensors were placed at points with vehicular activity, enabling them to record passing vehicles for model training. On the other hand, the professional sensory equipment used consisted of high-end seismographs, which are used to record seismic events. The key purpose is to convey the amalgamated deep neural network with the capability to effectively discern earthquakes in signals from low-cost sensors that are contaminated with vehicular noise, thereby avoiding false positives caused by vehicles. The proposed low-cost sensors could be used in bulk and placed in different areas so they can record an upcoming

seismic event. This could benefit researchers and give them the ability to record the events from different sensors and extract valuable information. In addition, the low-cost sensors are easy to maintain or replace, given their affordability in comparison to professional equipment, and could be placed near roads for easier access to them. Finally, the model we propose could be very useful when it comes to earthquake recognition, as it has the ability to recognize the seismic event from a passing vehicle; our model is trained to discern the difference between them. This model supports our proposal of placing the low-cost sensors near roads for easier access, as passing vehicles will not affect the recognition process of the model. A bird's-eye view of the key pillars of this work is presented in Figure 1.

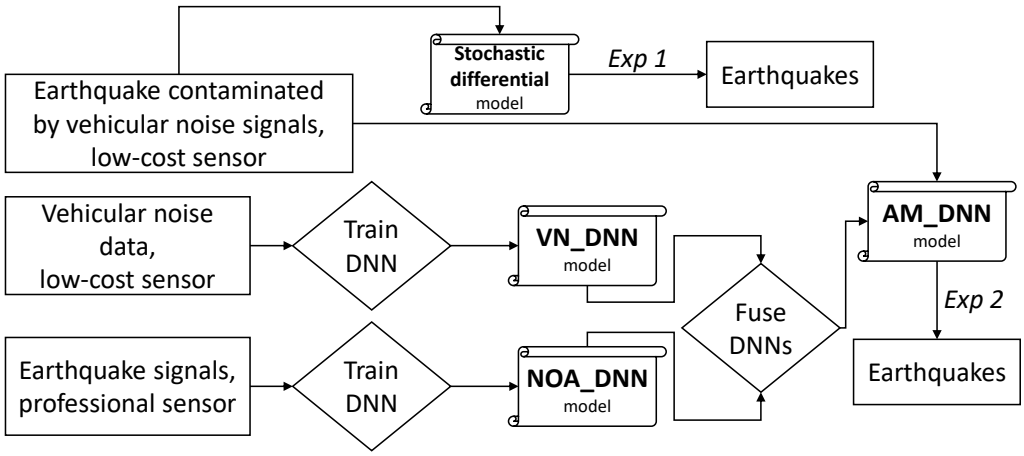

Figure 1. Architectural diagram of the proposed methodology.

The online availability of the programming code and data in the scientific research offers numerous benefits, i.e., it improves transparency, collaboration, and accountability by enabling independent verification of the findings. To this end, all data and code of this work are available online (https://github.com/LeonidasAgathos/Identifying_earthquakes_in_low_cost_sensors_signals_contaminated_with_vehicular_noise, accessed on 13 September 2023).

3.1. A Stochastic Approach

A very common scenario used to detect earthquake events involves signals from low-cost sensors, placed in several areas near roads and a methodology that allows the identification of earthquake events in such data. In our context, a seismic event is defined as the transition from a state of pure noise to a seismic signal. The current methodology we use to detect earthquake events, as presented in [27], is a stochastic differential model that employs a sliding window technique on the time-series data of the sensor. This window is incrementally moved through the dataset at predefined intervals. Within each window, the data undergo a transformation process, the variance function of the transformed data is computed, and its shape is assessed in relation to a power law distribution.

If the shape of the variance function closely aligns with the characteristics of a power law distribution, it is postulated that this is indicative of the data window predominantly representing noise. In this case, the model proceeds to the next window. However, when the shape of the variance function diverges significantly from the expected power law shape, the algorithm terminates and signals the detection of a seismic event.

In our case, we applied this algorithm to a dataset obtained from our low-cost seismic monitoring system. The successful detection of a seismic event was determined based on the algorithm's ability to identify an outbreak near the initiation of the seismic event, while not focusing on the detection of the whole event.

It is worth noting that, although this method demonstrated competency in identifying seismic events in its original presentation, it faced challenges in distinguishing between

seismic events and events triggered by external factors, such as passing vehicles. This limitation was discussed as a key factor in the interpretation of the results.

3.2. Low-Cost Seismic Sensory Equipment

The low-cost sensor mentioned in Section 3.1 was created in the CMODLab of Ionian University, Corfu, Greece; it consists of low-cost hardware and a data logger system. Originally, it was created for detecting seismic events and was placed in various areas, mostly near traffic roads. This placement is a part of the low-cost concept, so the sensors are easily accessible and replaceable in case of malfunction.

The system employs a 3-axis geophone, operating at 4.5 Hz and 380 Ohm, serving as the main data logger to record signals. A sampling rate of 225 Hz is archived from the data logger and an accurate timestamp is added to each sample from a precise real-time clock circuit. This clock is checked and corrected every hour by using internet information. The recorded data are stored internally at the system in 5 min chunks (coinciding with files); subsequently, these are transmitted to the database server every 5 min by using internet connectivity.

The low-cost system consists of low-cost hardware and open-source software. The system uses Raspberry Pi 3 B+, which is a credit-card-sized microcomputer board (see Figure 2). In addition, the system uses an analog-to-digital board with 24-bit high speed (ADS1256) precision, specialized for interconnection with the microcomputer and the real-time clock circuit breakout board DS3231. To support the system's energy needs, it uses a step-down converter with +5Volt power, up to 3 Amps. The system is also fitted with a solar panel, battery, and a solar charger controller to make the system autonomous. Moreover, the system utilized additional accessories, such as a USB GSM–GPRS 4G modem to support the system's internet connection, a 3-axis geophone sensor with a cutoff frequency *fc* set at 4.5 Hz, and a micro-SD 32 GB memory card that functions as the hard disk of the microcomputer board and stores all the necessary software needed to support the system (e.g., Python, data, etc.). The operating system used in the sensor is a Linux-based operational system for Raspberry Pi. We also used Python and C++ to write and execute scripts, depending on the needs of the task. The microcomputer board Raspberry Pi 3 B+, is the heart of the data logger system, and was selected for its high adaptability to integrate with various additional boards, like (I2c Bus, UART, SPI, GPIO, etc.), and it provides the data logger with multiple capabilities.

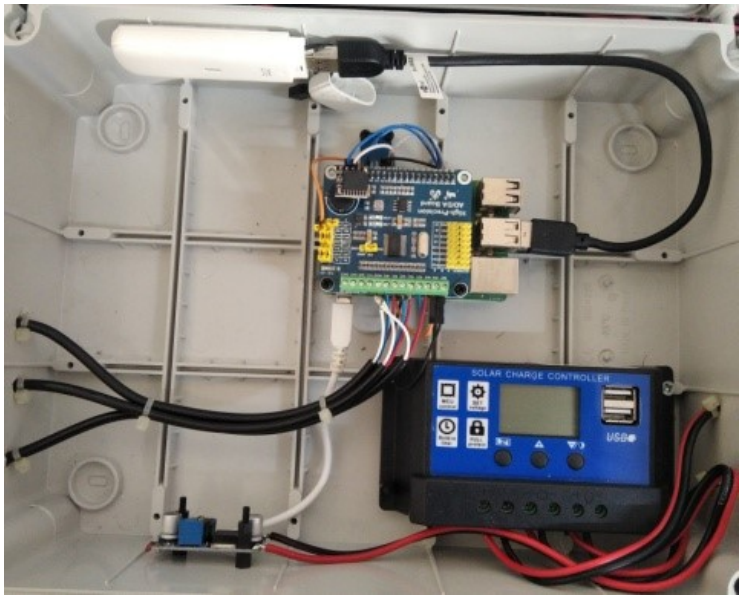

Figure 2. Full data logger system setup with housing in a plastic waterproof IP67 box.

In addition, as mentioned above, a 24-bit A/D high-speed analog-to-digital precision board (ADS1256), as shown in Figure 3, is connected to the microcomputer board using the 40-pin GPIO connector.

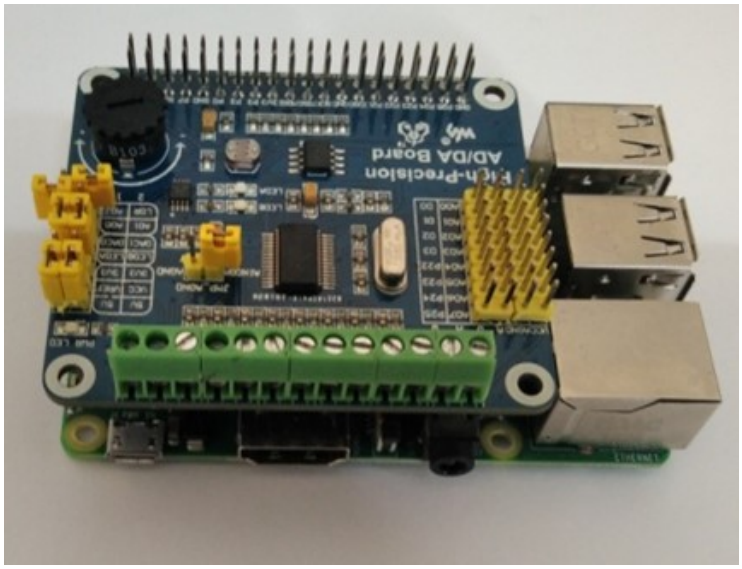

Figure 3. Raspberry Pi 3 B+ in combination with the 24 bit high precision A/D board—ADS1256.

The A/D board has 24 bits of accuracy and has a quantization error 1/2 LSB of 2^{24}/max input voltage. It can be adjusted to operate with a max input voltage of 3.3 or 5 volts. In our case, the max input voltage is 5 volts, which means we have a quantization error of 1/2 of 5 volts/2^{24}. The board has 8 analog inputs, which can work in a simple mode (8 input channels) or in a differential mode (4 input channels), similar to our data logger. It can accommodate sample rates of up to 30K samples per second (in a single channel—simple mode) and features an internal amplifier with an amplification factor of up to 64. In the proposed system, it uses a sampling rate of 3750 Hz and an amplification factor of 64. As per the datasheet (https://www.ti.com/lit/ds/symlink/ads1256.pdf, accessed on 13 September 2023) of the ADS1256, the noise level is up to 6 bits, while the effective number of bits (ENOB) with the buffer off is essentially the remaining 18 bits. The output of the A/D card is given in counts. According to the amplification factor of 64, the differential max input voltage that can be measured from the A/D card cannot be higher than ±78.125 mV. Each count has a value of $2^{24} \pm -78.125$ mV, so that means that each count has a value of 0.000009312 mVolts (minimum count step).

The system's 3-axis 4.5 Hz geophone, as shown in Figure 4, is a SEIS (https://www.seis-tech.com/4-5hz-3c-geophone/, accessed on 13 September 2023) 4.5 Hz 3C geophone, and it is connected to our data logger via inputs of the A/D high-speed analog-to-digital precision board. The geophone is connected in differential mode (A0–A1 input for X-Axis, A2–A3 input for Y-Axis, A4–A5 input for Z-Axis, while A6–A7 is not used). The sensitivity of the geophone for each axis, as per the datasheet (https://www.seis-tech.com/wp-content/uploads/2022/01/3c-geophone-4.5hz.pdf, accessed on 13 September 2023), is about 28.8 Volt/m/s (in open circuit) or 0.0288 Volt/mm/s. Finally, according to the maximum input voltage of the A/D card and the geophone output voltage, we can see that our system has an area to collect the ground velocity data of almost ±2.71 mm/s.

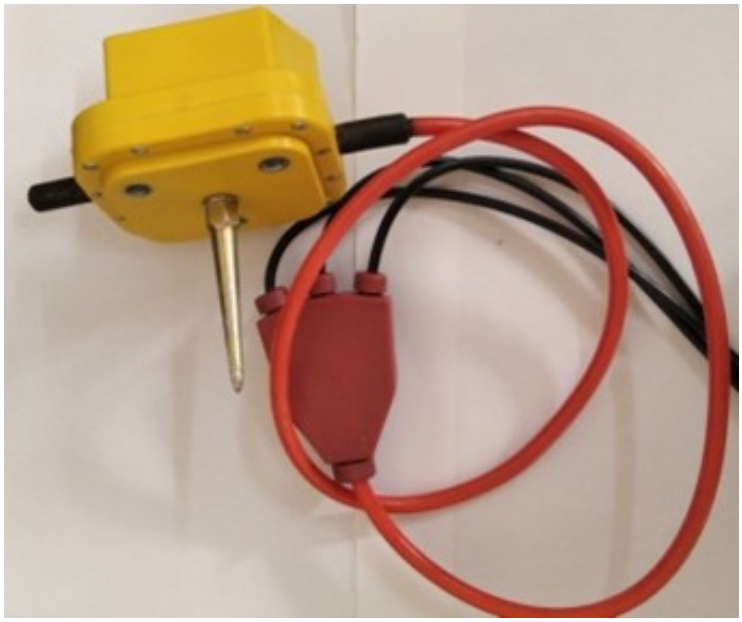

Figure 4. Three-axis geophone with cutoff frequency fc = 4.5 Hz.

3.3. Vehicular Noise

The acquisition of data originating from vehicular noise, utilizing the aforementioned geophone system in Section 3.2, constituted a crucial phase of our experiments. The key requirement was to identify a location to place the sensor on a major road with a gap close to the road for the positioning of the sensor. Additionally, in order to be able to monitor the collection process, the location had to be opportune for human operators tasked with recording the passing vehicles so that we could confirm the ground truth and align it with the signals captured by the geophone. To fulfill all these requirements, we selected a frequently accessed road in close proximity to our laboratory in the Garitsa area in Corfu, Greece. For all the above constraints, we collected the signals of passing cars via the geophone of the sensor, along with the audio recording, in order to assist in the labeling phase of ground truth later on. Using these recordings, we created an annotated dataset containing vehicular noise. In the post-collection process, to create the final dataset for vehicular noise, we cleaned and labeled the data by selecting the most representative axes of the geophone data containing the records from the movement of the ground. For the labeling process, the timestamps were labeled manually using the synchronous audio mentioned previously, collected concurrently with the geophone data, each time a vehicle passed. In order to further support the reproducibility of our work, the data of this dataset are available online (https://drive.google.com/drive/folders/1_H72gqp2ObBizB_YHRI0u53yTSHRiLqc?usp=drive_link, accessed on 13 September 2023).

3.4. Ground Truth Earthquake Dataset

After collecting the vehicular dataset, we also had to collect a dataset about earthquake events that would act as the ground truth. To complete this task, a data pipeline was created using the Obspy framework [28], which extracts waveform events from the European Integrated Data Archive in the National Observatory of Athens (NOA) [29] and saves the value of data and the timestamp of each signal wave in raw format. The seismic events were recorded and downloaded from the station *VLS* in Valsamata, Kefalonia, Greece, which is part of the NOA network. After collecting data for 773 seismic events from NOA for the VLS station, data cleaning was applied. All earthquakes were visually inspected for anomalies during their recording process. Earthquake signals that displayed irregular patterns in their recording before or after the main event were discarded. Figure 5 shows examples of regular and irregular earthquake signals. After the above phase, a total of 503 seismic events remained in our training dataset. To label the data and find the timestamps of the

start time and end time of each event, we used the STA/LTA Z-Detect [30] algorithm. This task was conducted using Obspy, which also features libraries for this task. Finally, and in preparation for feeding these data to the neural network, we created data frames for all the seismic events with their original values, and their values normalized in the range of $(-1, 1)$. In order to further support the reproducibility of our work, the data of this dataset are available online (https://drive.google.com/drive/folders/1AgB4aC3yI7axPM9Jp4RhvkwORBjOK5ST?usp=drive_link, accessed on 13 September 2023).

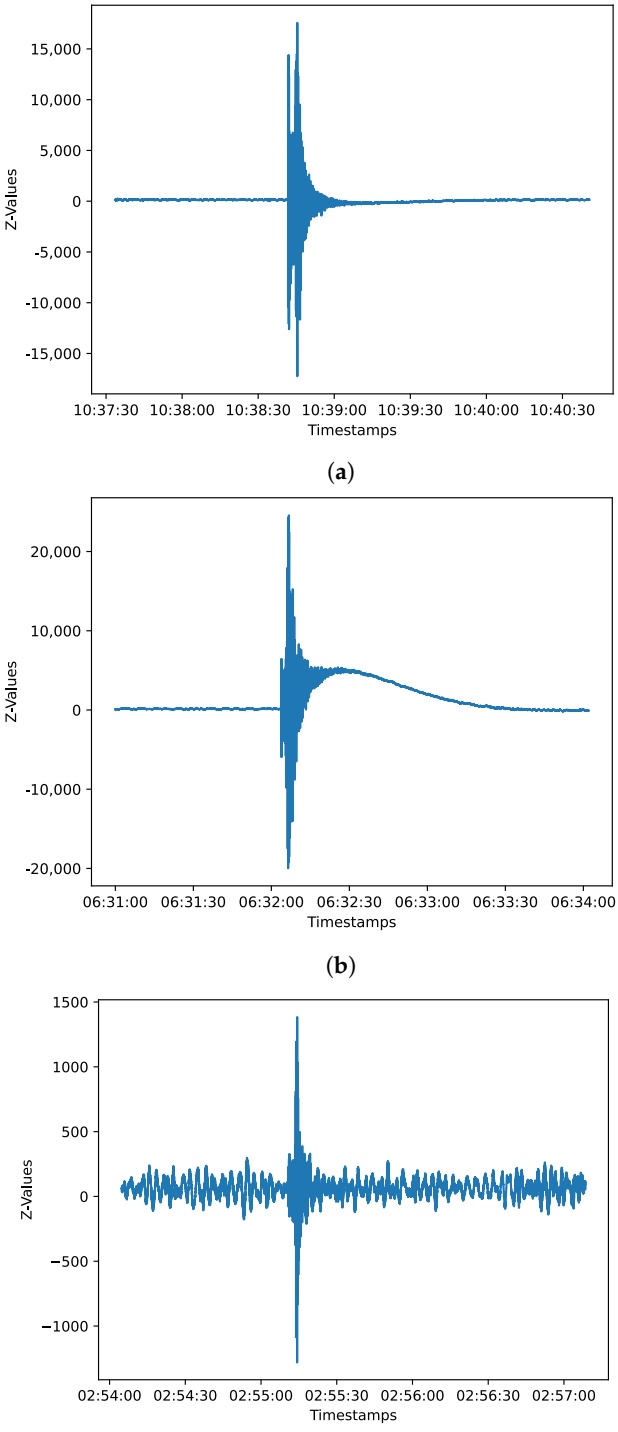

Figure 5. (**a**) Regular earthquake signal; (**b**) irregular earthquake signal; (**c**) irregular earthquake signal.

3.5. Training Process and Creation of DNNs

In order to train the models discussed herein, the following tasks were performed:

- Data preparation included several format conversion tasks aimed at converting the data into a proper form;
- Data normalization, wherein data were linearly normalized in the range of $[-1, 1]$;
- Class imbalance handling, dealt with the imbalance of the dataset using the NearMiss method [31];
- Train–test split, where the available data were split in training and testing by means of a generic approach of an 80–20% split, so we could use enough data to train the models.

In order to create the classification models based on each of the aforementioned datasets (Sections 3.3 and 3.4), we utilized the TensorFlow [32] and Keras [33] frameworks. Both these frameworks are renowned for their high performance. In detail, the used classification model is a long short-term memory model [34], which contains five layers for training and validation:

- A hidden LSTM (long short-term memory neural network) [35] layer with 64 units and a *return_sequences* = *True* parameter, which returns the full sequence of outputs for each input sequence and allows stacking additional recurrent layers;
- A 'flatten' layer [2], which flattens the 3D output from the LSTM layer into a 2D tensor; this is typically done to connect the LSTM layer to a standard feed-forward neural network;
- Two dense layers [36]: The first layer comprises 32 units and the second of 16 units, which are fully connected, and each neuron is connected to every neuron in the previous layer. Both dense layers use the ReLU (rectified linear unit) [37] as activation functions;
- An output layer, which is also a dense layer that represents the output of the model. The activation function used in this case is the sigmoid function [38], which outputs a probability score between 0 and 1.

3.6. Two-Fold Synchronized Dataset

The final seismic signal dataset consists of time-series data collected from our low-cost system, described in Section 3.2, strategically placed in a region prone to seismic activity. This dataset primarily comprises seismic signals associated with earthquake events. These signals exhibit a wide spectrum of characteristics, encompassing various magnitudes and frequencies. Of particular significance is the inclusion of ambient noise originating from passing vehicular traffic. This environmental noise component, stemming from the dataset's proximity to a roadway, introduces a distinctive dimension to our dataset. While seismic signals provide insights into genuine ground motion events, the presence of vehicular noise poses a challenge that reflects real-world scenarios and presents a challenge to our machine learning model.

To ensure the dataset's reliability, we cross-referenced our recorded signals with data from established seismographs from NOA, known for their accuracy and trustworthiness. The purpose of this dataset is to test both the stochastic and amalgamated models, in order to verify their ability to distinguish between earthquake signals against vehicular noise signals. The dataset was recorded and saved into 84 distinct csv files, which then were visually inspected, and each data point was labeled either as an earthquake or noise (irrespective of being vehicular or otherwise). Later, the same procedure as with the previous datasets was performed to prepare it for the testing phase. In order to further support the reproducibility of our work, the data of this dataset are available online (https://drive.google.com/drive/folders/16uKG9eq1kkf9Xpk39Tt96HHBtZ06N6qq?usp=drive_link, accessed on 13 September 2023).

4. Experimental Evaluation

This section details the setup that was used for performing the experimental evaluation of the proposed methodology, as well as the results received.

4.1. Experimental Setup

For the experimental part of our research, we utilized the datasets previously described: the dataset for vehicular noise (Section 3.3), the ground truth earthquake dataset (Section 3.4), as well as the two-fold synchronized dataset (Section 3.6). The former aims at providing information on vehicular noise, as perceived by the proposed low-cost sensory equipment. The second dataset aims to act as the ground truth point of reference, given its provenance from the European Integrated Data Archive in the National Observatory of Athens, and the fact that the sensory equipment used to collect these data is of very high accuracy. The latter dataset is the combined and synchronized dataset of seismic data from the low-cost seismic sensor in addition to seismic data from the professional seismic sensor, while both sensors were in very close proximity. All these datasets are also available (https://github.com/LeonidasAgathos/Identifying_earthquakes_in_low_cost_sensors_signals_contaminated_with_vehicular_noise/blob/main/Data_Availability, accessed on 13 September 2023). For the training part of the process, as extensively discussed in Section 3.5, the TensorFlow and Keras frameworks were utilized to create the classification models. The creation of DNNs was based on the LSTM classification model (see Section 3.5 for more details) using five layers for training and validation. The hardware configuration used was a computer with an i7-9700k CPU, featuring 8 cores and 16 gigabytes of RAM, along with an MSI GTX 1660 Ti GPU, equipped with a 6-gigabyte memory card. In order to evaluate the results received, the metrics used herein were accuracy, precision, recall, F1 score, and the area under the curve (AUC) [39], as per the following equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall or Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1 score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

where TN, FN, FP, and TP are true negative, false negative, false positive, and true positive, respectively.

4.2. Training Model 1

In this training model, we trained a DNN (https://github.com/LeonidasAgathos/Identifying_earthquakes_in_low_cost_sensors_signals_contaminated_with_vehicular_noise/blob/main/Model_for_Car.ipynb, accessed on 13 September 2023) to identify vehicular signals so that the model can predict when we have a signal from a passing vehicle or pure noise. To achieve that, we had to normalize the data in the range (1, −1) and additionally balance the data. The balancing was done by undersampling the majority class (noise class) using the method NearMiss (version1), so we have the same amount of data on both classes. After the pre-processing of the data, we created the DNN, which contained four layers (as per Section 3.5). We used one LSTM layer followed by a 'flatten' layer and three dense layers. LSTM layers are widely used for time series predictions as they can learn patterns and correlations within the time series, crucial for earthquake detection. The other layers are simple ones that assist in transforming the data. The data were allocated with 80% for training and 20% for testing. The results of the training are shown in Table 1 while the ROC curve is shown in Figure 6.

Table 1. Performance metrics for training model 1.

Accuracy	Precision	Recall	F1 Score	AUC Curve
68%	77%	52%	69%	73%

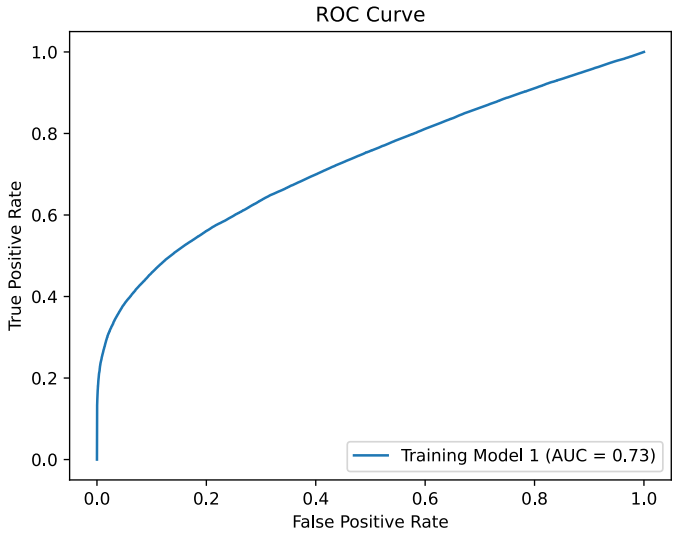

Figure 6. AUC for Model 1.

The results, despite being promising, are far from optimal because of the complexity of the data we used for training. Also, signals were labeled manually using the synchronized audio files as ground truth, which means that minor discrepancies between the audio and the file may have occurred.

4.3. Training Model 2

The second model (https://github.com/LeonidasAgathos/Identifying_earthquakes_in_low_cost_sensors_signals_contaminated_with_vehicular_noise/blob/main/Model_for_Eartquake.ipynb, accessed on 13 September 2023) was trained to identify earthquake events. To perform this task, the NOA data were used, which, as mentioned above in Section 3.4, includes 502 seismic events. After collecting the data, during the pre-processing phase, normalization was applied in the (−1, 1) range. After that, we performed the balancing of the data and the creation of the DNN. As mentioned before, the balancing method we used was NearMiss (version 1) and the DNN contained the same layers as the prior training model 1 (Section 4.2) for the same reasons mentioned above. In this model, samples were classified as either noise or earthquake. Table 2 shows the training and test results while the ROC curve is shown in Figure 7.

Table 2. Performance metrics for training model 2.

Accuracy	Precision	Recall	F1 Score	AUC Curve
75%	83%	63%	72%	82%

The training was conducted in 10 epochs, selected to avoid over-fitting, and lasted approximately 2 min. The results received from the training and testing phase were better than the previous model but again far from optimal.

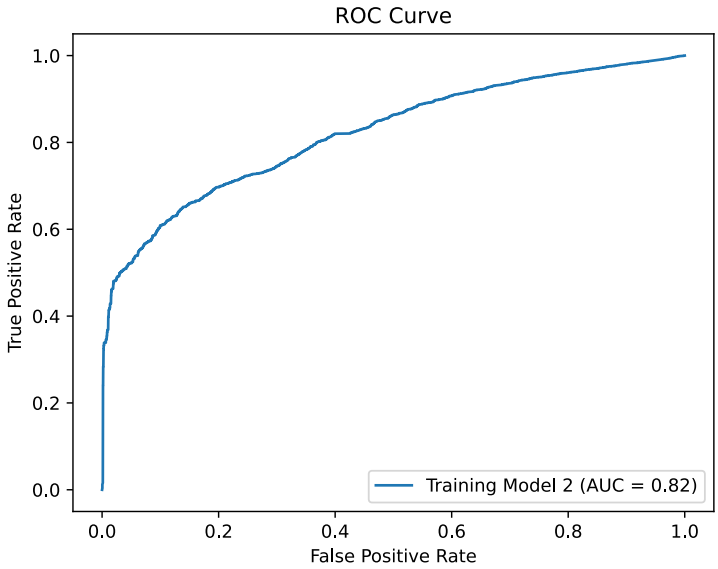

Figure 7. AUC for Model 2.

4.4. Experiment 1

The first experiment in this study employed the stochastic differential model, as described in Section 3.1. The primary objective of this model was to identify the onset of seismic events within an arbitrary time-series dataset. To do that, the two-fold synchronized dataset was fed to the stochastic differential model. Then, we extracted the actual starting point of every earthquake event and compared it to the true labels held on the two-fold synchronized dataset files. Finally, we had to evaluate the results and extract the metrics of the results.

The results obtained from this experiment can be seen in Table 3 and Figure 8. The metrics received from the first experiment show us a moderate performance of the model; the accuracy and precision reached 46% while the F1 score was 63%. Also, the AUC curve (shown in Figure 9) was 50% and the recall was at 100%.

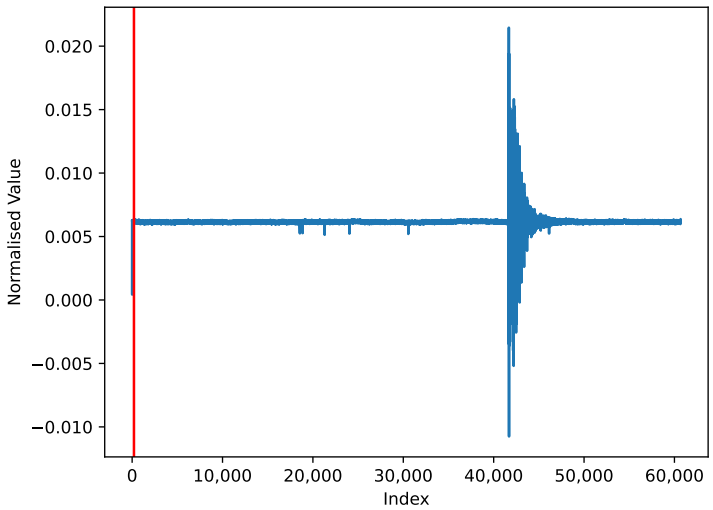

Figure 8. *Cont.*

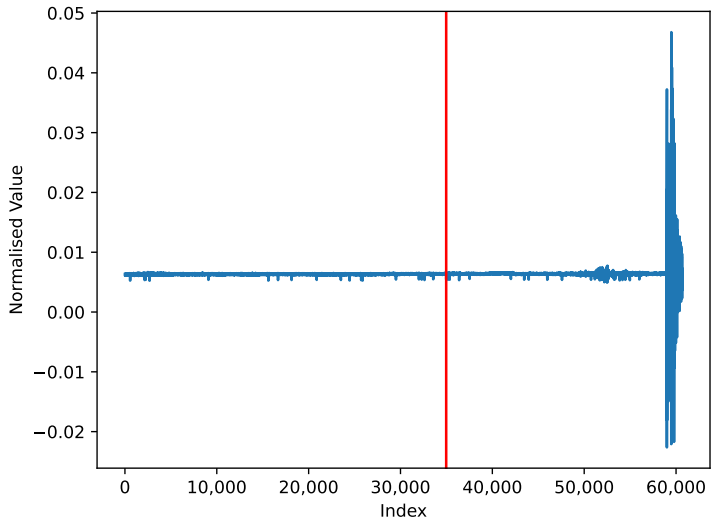

Figure 8. Results of the stochastic algorithm's performance.

Table 3. Performance metrics for experiment 1.

Accuracy	Precision	Recall	F1 Score	AUC Curve
46%	46%	100%	63%	50%

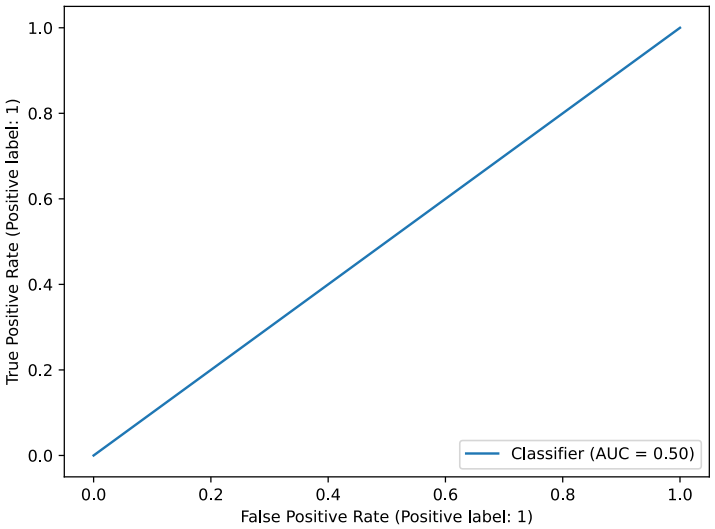

Figure 9. AUC for the stochastic model.

4.5. Experiment 2

In our second experiment, we created an amalgamated DNN model (https://github.com/LeonidasAgathos/Identifying_earthquakes_in_low_cost_sensors_signals_contaminated_with_vehicular_noise/blob/main/Model_Concatenation.ipynb, accessed on 13 September 2023), which includes the two models from Sections 4.2 and 4.3. The purpose of this experiment is for our model to be able to classify the starting point of an earthquake event without confusion from car signals and to perform better than the current methodology used by the stochastic model. This was achieved by combining the two different trained models and concatenating them, thus resulting in a single tensor that is the concatenation of all inputs. Initially, the pre-trained models of noise–cars and noise–earthquakes were loaded, and we applied the concatenation process. After that, we had to load the two-fold synchronized dataset and pre-process the data, which pertained to dropping the unnecessary columns

and normalizing the data. With our two models in the same shape and the two-fold synchronized dataset in the proper form, our final model was ready to classify the two-fold synchronized dataset. We set a dense layer with the activation function "sigmoid", as we needed to differentiate between noise and earthquakes. We considered car signals and noise in the same class, as the task was to find the earthquake events and separate them from any noise signals. To compare the results, we had to classify the two-fold synchronized dataset based on the final model we created. After that, we had to find the first point of every file in the two-fold synchronized dataset, which was classified as an earthquake in order to determine the starting point of the event. This would also be the point where we would assess proper metrics and compare the two experiments. We saved the classified labels along with the true labels, so we could derive the metrics from this experiment. The results of the metrics can be seen in Table 4.

Table 4. Performance metrics for experiment 2.

Accuracy	Precision	Recall	F1 Score	AUC Curve
78%	78%	100%	88%	51%

As we can see, the accuracy is 78%, which means that we see a big improvement when compared to the results of the experiment in Section 4.4, and our model was able to classify the start of the event, in a much more efficient way. Also, the precision of this experiment is 78%, which means our model performed well in finding the actual start of the earthquake events (the positive class). The recall was found to be at 100% due to the one class we have in the predicted labels (we only kept the positive class, which indicates the earthquake). Finally, the F1 score was found to be at 88%, which is likely the most important metric to consider in this experiment, as our target was to classify as many true positives as possible and the AUC curve (shown in Figure 10) was at 51%, which is probably caused by the lack of the negative class in the given dataset.

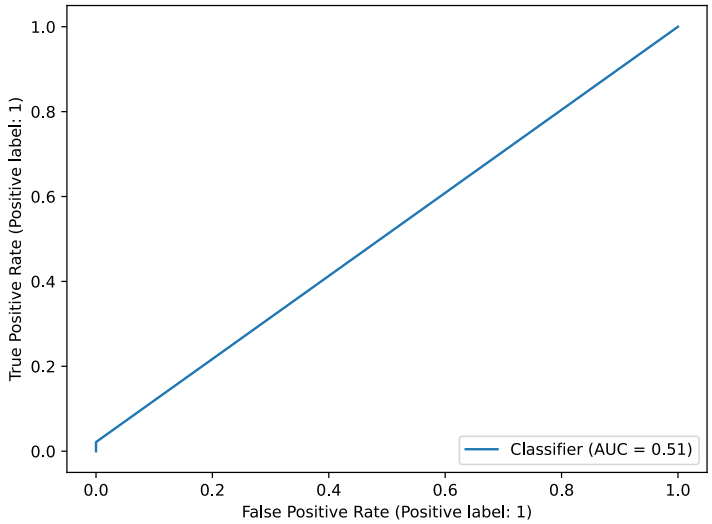

Figure 10. AUC for the proposed model.

In summary, the results, as illustrated in Figure 11, firmly validate the efficacy of our proposed methodology when compared to the stochastic approach. Our proposed model exhibits notably superior performance, emphasizing its potential in practical applications. Specifically, our proposed model achieves comparable results in the recall metric when contrasted with the stochastic model. However, it significantly outperforms the latter by achieving a precision score exceeding 30%, underscoring its capacity to accurately identify positive instances (true positives) while mitigating the occurrence of false positive errors. Moreover, our model attains a substantially improved F1 score, surpassing the stochastic

model by more than 20%. This superior F1 score attests to our model's ability to strike an optimal balance between precision and recall, making it a promising choice for various real-world scenarios.

As shown in Figure 12, we can see how our model performed in real-world scenarios. The red color presents the space between the first point predicted as the earthquake to the last point predicted as the earthquake. The two vertical lines show the actual start and end of the earthquake (ground truth). When it comes to the start of the event, our model classifies the starting point exactly at the onset. As we can also observe, the model classifies the whole earthquake event, not only the starting point. The ending part of an earthquake is always a complex task so the model performs decently (regarding classification) on that as well. When comparing these results with the stochastic methodology (Section 4.4), we can see a big improvement in finding the earthquake events, not only on the metrics, but also in the actual usage of a model like this, which detects an earthquake event.

While our cost-effective system, built on open hardware and software, offers an attractive alternative to pricier traditional seismographs, it comes with inherent limitations in the sampling rate, bit precision, and amplification. Our dataset, a combination of high-quality and low-cost recordings, presents an imbalance issue due to the significantly smaller proportion of low-cost data. This imbalance affects our model's outcomes. Additionally, the constraints tied to our DNN's architecture were expected, given its off-the-shelf nature. However, this architecture can be enhanced through increased parameterization.

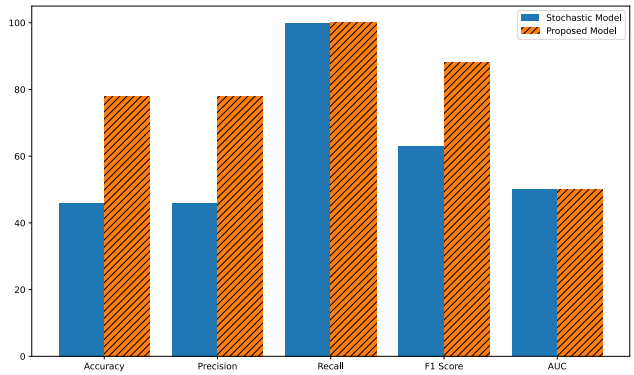

Figure 11. Comparison of metric scores.

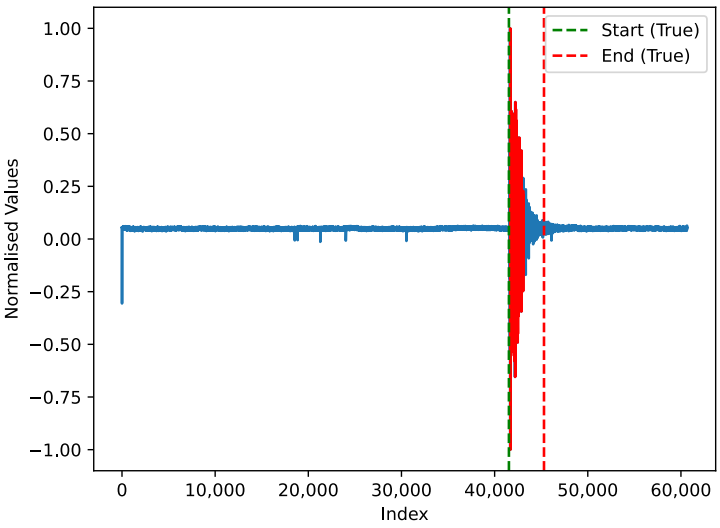

Figure 12. *Cont.*

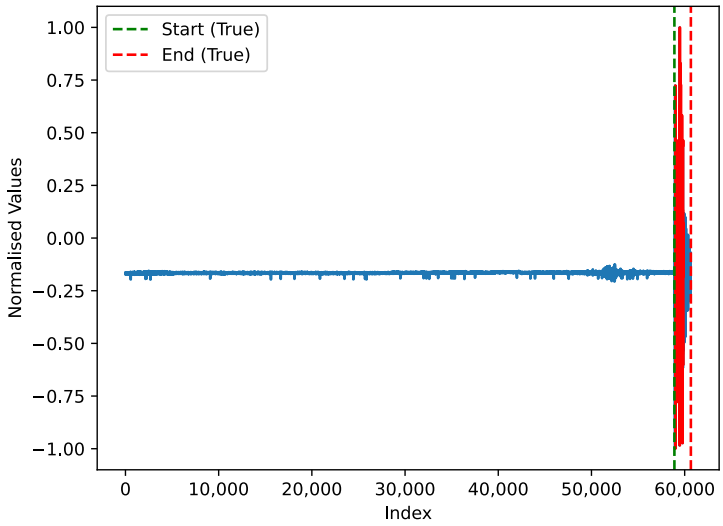

Figure 12. Results of the amalgamated DNN model's performance.

5. Conclusions

This paper underscores the profound significance of earthquake monitoring for the purposes of disaster management, public safety, and scientific research. Moreover, the pivotal role of sensor networks in amassing seismic data is presented, forming the cornerstone of current earthquake detection capabilities. In that context, the advent of low-cost sensors has unlocked the potential for their extensive deployment in the field, despite their susceptibility to vehicular noise pollution. Low-cost seismic sensors present a formidable challenge, necessitating innovative solutions for accurate signal extraction. Since the importance of precise earthquake identification cannot be overstated, as it holds the key to timely alerts and informed decision-making, herein, we propose the use of an amalgamated deep neural network (DNN) composed of (i) a DNN trained on earthquake signals from professional sensory equipment, as well as (ii) a DNN trained on vehicular signals from low-cost sensors, for the purpose of earthquake identification in signals from low-cost sensors contaminated with vehicular noise.

Our proposal includes a detailed presentation of a low-cost seismic sensory equipment, which is designed to be approximately two orders of magnitude less expensive than typical professional seismic measurement equipment. Still, the low-cost seismic sensory equipment was shown to be prone to vehicular noise contamination. Accordingly, the proposed amalgamated deep neural network underwent evaluation through experimentation and has manifested significant performance improvements compared to a generic stochastic differential model. The superiority of the proposed methodology addresses the need of effectiveness, as it identifies both the beginning and the end of a seismic event, as well as the need of efficiency, as indicated by the performance measures. Future plans will include customizing the generic DNNs deployed in this study for the task at hand, in order to address the necessities of the work's scenario and achieve even higher efficiency in the end identification process. Moreover, we plan to expand the two training datasets (vehicular noise and ground truth from professional seismometers) provided herein to ensure more generality and to better train the DNN, accordingly obtaining more general results. Finally, we plan to extend our low-cost network of sensors in locations near the professional seismometers in order to enhance the two-fold synchronized dataset and to test even more diverse scenarios.

Author Contributions: Conceptualization, L.A., A.A., N.A., I.V., I.K. and M.A.; Methodology, L.A., A.A., N.A., I.V., I.K. and M.A.; Software, L.A., A.A., N.A., I.V., I.K. and M.A.; Validation, L.A., A.A., N.A., I.V., I.K. and M.A.; Formal analysis, L.A., A.A., N.A., I.V., I.K. and M.A.; Investigation, L.A., A.A., N.A., I.V., I.K. and M.A.; Resources, L.A., A.A., N.A., I.V., I.K. and M.A.; Data curation, L.A., A.A., N.A., I.V., I.K. and M.A.; Writing—original draft, L.A., A.A., N.A., I.V., I.K. and M.A.; Writing—review & editing, L.A., A.A., N.A., I.V., I.K. and M.A.; Visualization, L.A., A.A., N.A., I.V., I.K. and M.A.; Supervision, L.A., A.A., N.A., I.V., I.K. and M.A.; Project administration, L.A., A.A., N.A., I.V., I.K. and M.A.; Funding acquisition, L.A., A.A., N.A., I.V., I.K. and M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lim, J.; Jung, S.; JeGal, C.; Jung, G.; Yoo, J.H.; Gahm, J.K.; Song, G. LEQNet: Light Earthquake Deep Neural Network for Earthquake Detection and Phase Picking. *Front. Earth Sci.* **2022**, *10*, 848237. [CrossRef]
2. Ji, L.; Zou, Y.; He, K.; Zhu, B. Carbon futures price forecasting based with ARIMA-CNN-LSTM model. *Procedia Comput. Sci.* **2019**, *162*, 33–38. [CrossRef]
3. Murti, M.; Junior, R.; Najah, A.M.; Elshafie, A. Earthquake multi-classification detection based velocity and displacement data filtering using machine learning algorithms. *Sci. Rep.* **2022**, *12*, 21200. [CrossRef]
4. Mousavi, S.M.; Sheng, Y.; Zhu, W.; Beroza, G.C. STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI. *IEEE Access* **2019**, *7*, 179464–179476. [CrossRef]
5. Mao, F.; Khamis, K.; Krause, S.; Clark, J.; Hannah, D.M. Low-cost environmental sensor networks: Recent advances and future directions. *Front. Earth Sci.* **2019**, *7*, 221. [CrossRef]
6. D'Alessandro, A.; Scudero, S.; Vitale, G. A review of the capacitive MEMS for seismology. *Sensors* **2019**, *19*, 3093. [PubMed]
7. Arora, N.; Kumar, Y. Automatic vehicle detection system in Day and Night Mode: Challenges, applications and panoramic review. *Evol. Intell.* **2023**, *16*, 1077–1095. [CrossRef]
8. Mondal, B. Pandemic COVID-19, Reduced Usage of Public Transportation Systems and Urban Environmental Challenges: Few Evidences from India and West Bengal. In *Environmental Management and Sustainability in India: Case Studies from West Bengal*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 341–368.
9. Yu, Q.; Wang, C.; Li, J.; Xiong, R.; Pecht, M. Challenges and outlook for lithium-ion battery fault diagnosis methods from the laboratory to real world applications. *eTransportation* **2023**, *17*, 100254. [CrossRef]
10. Bhatia, M.; Ahanger, T.A.; Manocha, A. Artificial intelligence based real-time earthquake prediction. *Eng. Appl. Artif. Intell.* **2023**, *120*, 105856. [CrossRef]
11. Kassa, A.B.; Dugda, M.T.; Lin, Y.; Seifu, A. Earthquake Aftershocks Pattern Prediction. In Proceedings of the AGU Fall Meeting, Chicago, IL, USA, 12–16 December 2022; Volume 2022, p. S42C–0177. Available online: https://ui.adsabs.harvard.edu/abs/2022AGUFM.S42C0177D (accessed on 13 September 2023).
12. Shearer, P.M. *Introduction to Seismology: The Wave Equation and Body Waves*; Institute of Geophysics and Planetary Physics, Scipps Institution of Oceanography, University of California: San Diego, CA, USA, 2010; unpublished.
13. Udias, A.; Buforn, E. *Principles of Seismology*; Cambridge University Press: Cambridge, UK, 2017.
14. Kennett, B.L.N. *The Seismic Wavefield: Volume 1, Introduction and Theoretical Development*; Cambridge University Press: Cambridge, UK, 2001; Volume 1.
15. Hou, Y.; Jiao, R.; Yu, H. MEMS based geophones and seismometers. *Sens. Actuators Phys.* **2021**, *318*, 112498. [CrossRef]
16. Bullen, K.E.; Bolt, B.A. *An Introduction to the Theory of Seismology*; Cambridge University Press: Cambridge, UK, 1985.
17. Kumar, D.; Ahmed, I. Seismic Noise. In *Encyclopedia of Solid Earth Geophysics*; Gupta, H.K., Ed.; Springer International Publishing: Cham, Switzerland, 2021; pp. 1442–1447. [CrossRef]
18. Dou, S.; Lindsey, N.; Wagner, A.M.; Daley, T.M.; Freifeld, B.; Robertson, M.; Peterson, J.; Ulrich, C.; Martin, E.R.; Ajo-Franklin, J.B. Distributed acoustic sensing for seismic monitoring of the near surface: A traffic-noise interferometry case study. *Sci. Rep.* **2017**, *7*, 11620. [CrossRef] [PubMed]
19. Sun, L.; Qiu, X.; Wang, Y.; Wang, C. Seismic Periodic Noise Attenuation Based on Sparse Representation Using a Noise Dictionary. *Appl. Sci.* **2023**, *13*, 2835. [CrossRef]
20. Du, R.; Liu, W.; Fu, X.; Meng, L.; Liu, Z. Random noise attenuation via convolutional neural network in seismic datasets. *Alex. Eng. J.* **2022**, *61*, 9901–9909. [CrossRef]

21. Prasanna, R.; Chandrakumar, C.; Nandana, R.; Holden, C.; Punchihewa, A.; Becker, J.S.; Jeong, S.; Liyanage, N.; Ravishan, D.; Sampath, R.; et al. "Saving Precious Seconds"—A novel approach to implementing a low-cost earthquake early warning system with node-level detection and alert generation. *Informatics* **2022**, *9*, 25. [CrossRef]
22. Wu, Y.M.; Mittal, H. A review on the development of earthquake warning system using low-cost sensors in Taiwan. *Sensors* **2021**, *21*, 7649. [CrossRef]
23. Lee, J.; Khan, I.; Choi, S.; Kwon, Y.W. A smart iot device for detecting and responding to earthquakes. *Electronics* **2019**, *8*, 1546. [CrossRef]
24. Ahmad, A.B.; Saibi, H.; Belkacem, A.N.; Tsuji, T. Vehicle Auto-Classification Using Machine Learning Algorithms Based on Seismic Fingerprinting. *Computers* **2022**, *11*, 148. [CrossRef]
25. Nie, T.; Wang, S.; Wang, Y.; Tong, X.; Sun, F. An effective recognition of moving target seismic anomaly for security region based on deep bidirectional LSTM combined CNN. *Multimed. Tools Appl.* **2023**. [CrossRef]
26. Münchmeyer, J.; Woollam, J.; Rietbrock, A.; Tilmann, F.; Lange, D.; Bornstein, T.; Diehl, T.; Giunchi, C.; Haslinger, F.; Jozinović, D.; et al. Which Picker Fits My Data? A Quantitative Evaluation of Deep Learning Based Seismic Pickers. *J. Geophys. Res. Solid Earth* **2022**, *127*, e2021JB023499.
27. Avlonitis, M. On the problem of early detection of users interaction outbreaks via stochastic differential models. *Eng. Appl. Artif. Intell.* **2016**, *51*, 92–96.
28. Krischer, L.; Megies, T.; Barsch, R.; Beyreuther, M.; Lecocq, T.; Caudron, C.; Wassermann, J. ObsPy: A bridge for seismology into the scientific Python ecosystem. *Comput. Sci. Discov.* **2015**, *8*, 014003. [CrossRef]
29. Evangelidis, C.P.; Triantafyllis, N.; Samios, M.; Boukouras, K.; Kontakos, K.; Ktenidou, O.; Fountoulakis, I.; Kalogeras, I.; Melis, N.S.; Galanis, O.; et al. Seismic Waveform Data from Greece and Cyprus: Integration, Archival, and Open Access. *Seismol. Res. Lett.* **2021**, *92*, 1672–1684.
30. Choubik, Y.; Mahmoudi, A.; Himmi, M.; El Moudnib, L. STA/LTA trigger algorithm implementation on a seismological dataset using Hadoop MapReduce. *Iaes Int. J. Artif. Intell. (IJ-AI)* **2020**, *9*, 269. [CrossRef]
31. Mani, I.; Zhang, I. kNN approach to unbalanced data distributions: A case study involving information extraction. In Proceedings of the Workshop on Learning from Imbalanced Datasets, Washington, DC, USA, 21 August 2003; Volume 126, pp. 1–7.
32. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-Scale Machine Learning. In Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
33. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 13 September 2023).
34. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
35. Kowsher, M.; Tahabilder, A.; Islam Sanjid, M.Z.; Prottasha, N.J.; Uddin, M.S.; Hossain, M.A.; Kader Jilani, M.A. LSTM-ANN & BiLSTM-ANN: Hybrid deep learning models for enhanced classification accuracy. *Procedia Comput. Sci.* **2021**, *193*, 131–140.
36. Abualhaol, I.; Falcon, R.; Abielmona, R.; Petriu, E. Data-Driven Vessel Service Time Forecasting using Long Short-Term Memory Recurrent Neural Networks. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; Volume 12, pp. 2580–2590. [CrossRef]
37. Agarap, A.F. Deep Learning using Rectified Linear Units (ReLU). *arXiv* **2018**, arXiv:1803.08375.
38. Dubey, S.R.; Singh, S.K.; Chaudhuri, B.B. A Comprehensive Survey and Performance Analysis of Activation Functions in Deep Learning. *arXiv* **2021**, arXiv:2109.14545.
39. Yang, T.; Ying, Y. AUC Maximization in the Era of Big Data and AI: A Survey. *arXiv* **2022**, arXiv:2203.15046.