Breast cancer detection using a CNN: LightGBM approach

Idea for method

In the last ten years, Neural Network advances have been applied to mammography to help radiologists increase their efficiency and accuracy [1]. A wide variety of computer aided detection (CAD) approaches have been suggested, with one of the most cited in recent years the convolutional neural network with support vector machine (CNN: SVM approach) [2]. This article proposes a new CAD system with 5 elements: (1) image enhancement, (2) image segmentation, (3) feature extraction, (4) feature classification, and finally, (5) an evaluation for the classifier. In this CAD system, step 3 is done by the CNN because CNNs have been known to perform extremely well for extracting features [3], while step 4 is done with the help of an SVM as this is proven to be a good performing classification algorithm. In another study [4], another CAD was proposed, using a combination of a CNN with Random Forest (RF). The latter being an ensemble learning algorithm that constructs many decision trees during training. Although both SVM and RF can work well for classification tasks, a newer more efficient algorithm called Light Gradient Boosting Machine (LGBM), has been proven to outperform SVM and RF in a lot of classification tasks. LGBM is a highly efficient gradient-boosting decision tree that has been widely used by several winners of various machine-learning competitions and is already tested in the CNN: LGBM approach for bearing fault diagnosis [5]. However, it is not yet tested for detection of breast cancer using mammograms, and hence I would like to explore the 5-element CAD system mentioned earlier, but with the CNN: LGBM combination instead of the CNN: SVM combination.

Data

Possible data sources:
Mammography:

Table 1
Detailed information of the most commonly cited available mammography datasets. '-' means not available.

Dataset	Year	# Imgs	Format	View	Resln. (bit/pxl)	Pros	Cons	Select Publications
mini-MIAS [80]	2003	322	.PGM	MLO	8	The data can be accessed with Unix commands and is easy to retrieve.	Outdated film-screen mammograms and lacks more modern imaging sources like 3D-mammography. Limited to the MLO view.	[33,100,145]
DDSM [81]	1999	10480	.LJPEG	MLO, CC	12, 16	Commonly cited by the literature. Includes both the MLO and CC views.	Lacks an API and researchers need to install a special tool to retrieve images. Consists of outdated film mammography scans.	[33,47,48,120, 123,124,126, 146]
INBreast [82]	2011	410	. DICOM	MLO, CC	14	Both the MLO and CC views are available, and the images are widely cited.	The database is now restricted; Researchers must contact the authors directly for access.	[33,34,36,42, 111,117,120, 123]
BCDR [83,84]	2012	7315	.TIFF	MLO, CC	8; 14	Standard format Precise lesion locations. Includes BI-RADS density annotations and precise mass coordinates, as well as detailed segmentation outlines. Auxiliary patient data (prior surgery, lesion characteristics, biopsy status) is also available.	Limited to only 2D FFDM data. Limited in size.	[25,99,121]
BancoWeb LAPIMO [85]	2010	1400	.TIFF	MLO, CC	12	BI-RADs density labeled. Contains auxiliary information (patient age, scanner brand, hormone replacement therapy status). Standard image format.	Researchers need to wait for administrator approval. Limited in size.	[145]
VICTRE [86]	2018	217,913	. DICOM	MLO, CC	–	Contains precise of mammographic lesions. Entirely synthetic.	Entirely synthetic.	[147] [148]
OPTIMAM [88]	2020	>1 M	. DICOM	MLO, CC	12, 16	Extremely Large Dataset. Open-source API for easy image retrieval in Python.	Researchers need administrator approval for access. Data only comes from patients in England.	[35,132]

Figure 1: Caption

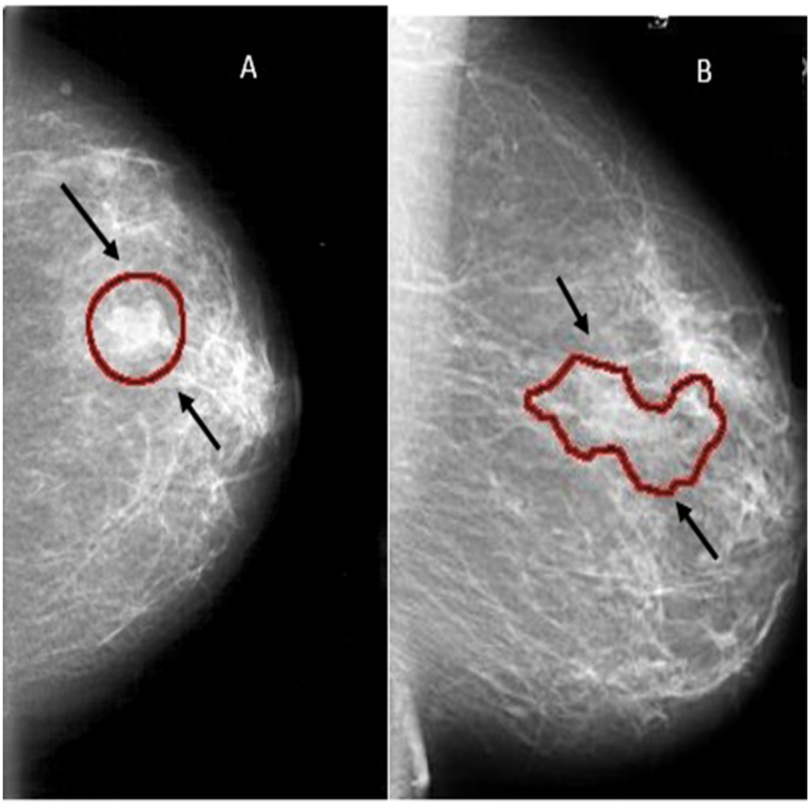

Figure 2: Caption

Histology:

1. The data from the UCI Repository's Wisconsin Breast Cancer Database (WBCD) have been used in this work. Malignant and benign types of cancer are the two major categories of information. The above dataset is used by most of the researchers for detecting cancer in the breast.

2. Widely used data: http://www.andrewjanowczyk.com/ use-case-6-invasive-ductal-carcinoma-idc-segmentation/

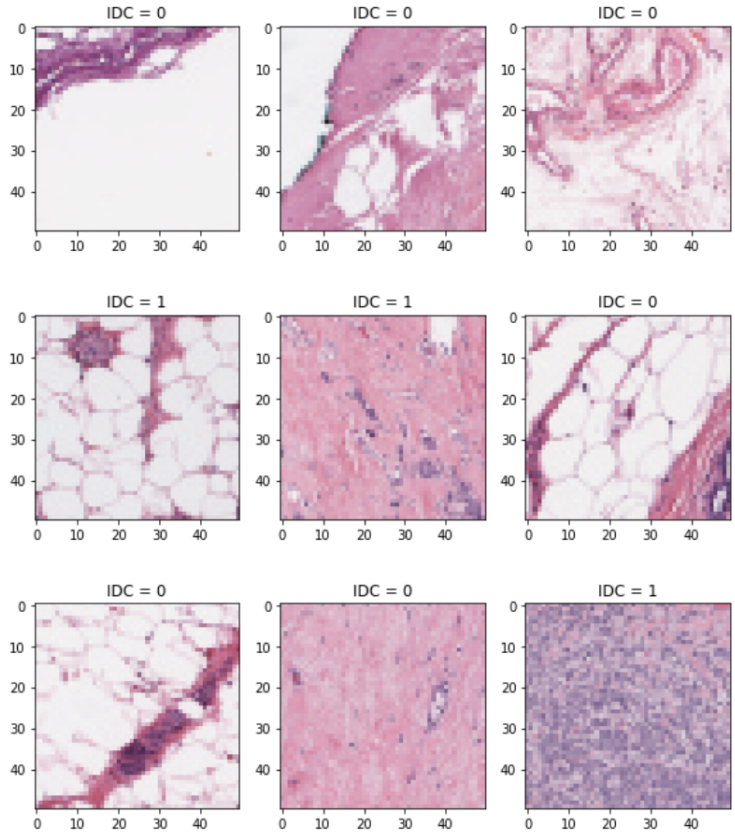

Figure 3: Caption

Articles

1. **Survey:** Leila Abdelrahman, Manal Al Ghamdi, Fernando Collado-Mesa, Mohamed Abdel Mottaleb, Convolutional neural networks for breast cancer detection in mammography: A survey, Computers in Biology and Medicine, Volume 131, 2021, 104248, ISSN 0010-4825

2. Dina A. Ragab et al. (2019) “Breast Cancer Detection Using Deep Convolutional Neural Networks and Support Vector Machines,” 7, p. 6201. doi: 10.7717/peerj.6201.

3. Meha Desai, Manan Shah, An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN), Clinical eHealth, Volume 4, 2021, Pages 1-11, ISSN 2588-9141

4. Almas Begum, V. Dhilip Kumar, Junaid Asghar, D. Hemalatha, G. Arulkumaran, "A Combined Deep CNN: LSTM with a Random Forest Approach for Breast Cancer Diagnosis", Complexity, vol. 2022, Article ID 9299621, 9 pages, 2022. https://doi.org/10.115

5. Xin Jia et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1043 022066

The M4 competition: Hybrid approach

Ideas for method

The M competition series was created to build a platform for researchers to test and compete forecasting models for a wide range of time series. The newest edition of the series is the M4 competition. This edition is based on 100,000 times series of different industries with different data frequencies. The organization proposed certain methods such as statistical approaches, ML approaches, and hybrid approaches, each having benchmark models to compare results with. The most notable result in the competition was a hybrid method proposed by (Smyl, 2020). His method used exponential smoothing (ES) in combination with recurrent neural networks (RNN). The hybrid forecasting approach had three main elements: (i) deseasonalization and adaptive normalization, (ii) generation of forecasts and (iii) ensembling. The ES was used for deseasonalization, and an long short term memory LSTM neural network was used for forecasting and ensembling. As the competition is based on a fast amount of time series, cross-learning for global subsets of time series and individual time series was applied, using ensembling of the two. My idea:

1. Implement a different method for deseasonalization of the time series. In the paper, it is mentioned that not all deseasonality algorithms are good preprocessors for RNN's. It is therefore suggested to use an algorithm that has integral parts that deal with the seasonality. I propose the following algorithms:

    - TBATS (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend, and Seasonal components): TBATS is an extension of the exponential smoothing model that can handle multiple seasonality components and other complexities, making it suitable for time series data with various seasonal patterns.
    - SARIMA: Seasonal autoregressive moving average model.

2. Replicate the method used for the competition. Since it is an advanced hybrid method with both statistical and ML approaches, I think I will learn a lot while having a sufficient enough methodology-level.

Data

The data were made available on 31st December, originally on the M4 website (www.m4.unic.ac.cy) and later via the M4comp2018 R package (Montero-Manso, Netto, & Tala-gala, 2018) and the M4 GitHub repository (www.github.com/M4Competition).

Table 1
Number of M4 series per data frequency and domain.

Time interval between successive observations	Micro	Industry	Macro	Finance	Demographic	Other	Total
Yearly	6,538	3,716	3,903	6,519	1,088	1,236	**23,000**
Quarterly	6,020	4,637	5,315	5,305	1,858	865	**24,000**
Monthly	10,975	10,017	10,016	10,987	5,728	277	**48,000**
Weekly	112	6	41	164	24	12	**359**
Daily	1,476	422	127	1,559	10	633	**4,227**
Hourly	0	0	0	0	0	414	**414**
Total	**25,121**	**18,798**	**19,402**	**24,534**	**8,708**	**3,437**	**100,000**

Figure 4: M4 competition data overview

Articles

1. (Smyl, 2020)

2. Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos, The M4 Competition: 100,000 time series and 61 forecasting methods, International Journal of Forecasting, Volume 36, Issue 1, 2020, Pages 54-74, ISSN 0169-2070

Notes

1. The ensembling method applied is one to implement cross-learning: it trains RNN's for global subsets of time series and individual time series. Since the amount of time series is very big, and training RNN's takes a long time, I do not know if I have sufficient computational power/time to use the same method.

2. From experience, I know that ES is very fast and efficient in comparison to the proposed deseasonalizing algorithms. Again, I might run into a computational/time issue.

The M4 competition: Pure ML approach

Idea for method

One of the findings of the competition is the lack of performances seen in the pure ML approaches, with the highest ranked ML approach (Trotta) averaging a rank of 24.33 over three different evaluation metrics (sMAPE, MASE, OWA), and the second highest ranked ML approach averaging 48,66 for those same metrics. The first approach implements a convolutional neural network (CNN), which are typically used for image based forecasting, and hence the approach applies a data transformation. The second highest uses a more simple method and implements a multiplayer perceptron network (MLP), also known as a feedforward neural network. Such methods can work well but typically have a hard time capturing patterns in sequential data like we are dealing with here. Therefore, I propose to implement a different kind of neural network called long short term memory neural networks (LSTM). Unlike traditional neural networks, LSTM incorporates feedback connections, allowing it to process entire sequences of data, not just individual data points. This makes it highly effective in understanding and predicting patterns in sequential data like time series.
Additionally, in [1], it is mentioned that the top performing approaches introduced information from multiple series (aggregated by data frequency) in order to decide on the most effective way of forecasting and/or selecting the weights for combining the various statistical/ML methods considered. However, the paper also mentioned that we should investigate such 'cross-learning' techniques to uncover its potential. That's why I suggest to build two separate models, one with cross-learning and one without to discover its potential for LSTMs.

Data

The data were made available on 31st December, originally on the M4 website (www.m4.unic.ac.cy) and later via the M4comp2018 R package (Montero-Manso, Netto, & Tala-gala, 2018) and the M4 GitHub repository (www.github.com/M4Competition)

Table 1
Number of M4 series per data frequency and domain.

Time interval between successive observations	Micro	Industry	Macro	Finance	Demographic	Other	Total
Yearly	6,538	3,716	3,903	6,519	1,088	1,236	**23,000**
Quarterly	6,020	4,637	5,315	5,305	1,858	865	**24,000**
Monthly	10,975	10,017	10,016	10,987	5,728	277	**48,000**
Weekly	112	6	41	164	24	12	**359**
Daily	1,476	422	127	1,559	10	633	**4,227**
Hourly	0	0	0	0	0	414	**414**
Total	**25,121**	**18,798**	**19,402**	**24,534**	**8,708**	**3,437**	**100,000**

Figure 5: M4 competition data overview

Articles

1. Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos, The M4 Competition: 100,000 time series and 61 forecasting methods, International Journal of Forecasting, Volume 36, Issue 1, 2020, Pages 54-74, ISSN 0169-2070

2. (Makridakis, Spiliotis, & Assi-makopoulos, 2018a)

Global air temperature forecast using a long short term memory neural network

Idea for method

Different research papers have reported that the climate will warm over the coming century, as a reaction to the changes in the anthropogenic emissions of CO_2 [1]. In a recent and greatly cited review paper on air temperature forecasting [2], it becomes clear that several papers have been published exploring machine learning as air temperature forecasting technique to help get a better picture of the warming climate. Among the published papers, some focused on global temperatures [3,4], and others on regional temperatures with different data frequencies [5,6]. Despite the researcher dealing with temporal data, only one paper proposed using a neural network suitable for this task: the long short term memory (LSTM) network. The key concepts of the LSTM have the ability to learn long-term dependencies by incorporating memory units while still having the general neural network feature: it can learn from very complex data patterns. Since temperature depends on a very complex system, the LSTM can be well suited for the task. However, the paper using the LSTM network only focused on hourly data of regional air temperature. Therefore, I propose to use an LSTM network and switch the focus to global temperatures mimicking the idea of a paper that used a multivariate neural network to predict the temperatures [4]. In particular, they assessed the CO_2 emissions inclusion in a nonlinear multivariate neural network, by means of data obtained from the annualised HadCrut3v (a data-set of land and ocean temperatures), and total carbon emissions from fossil fuels, between 1850 and recent years. Since HadCrut is already at version 5, I will not use version 3 but instead use the 5th version.

Data

The HadCRUT5 near surface temperature data set is produced by blending data from the CRUTEM5 surface air temperature dataset and the HadSST4 sea-surface temperature dataset. In figure (6), I plotted the monthly temperature data of the HadCRUT5 dataset. Note that the model would have features like CO_2 emission as well.

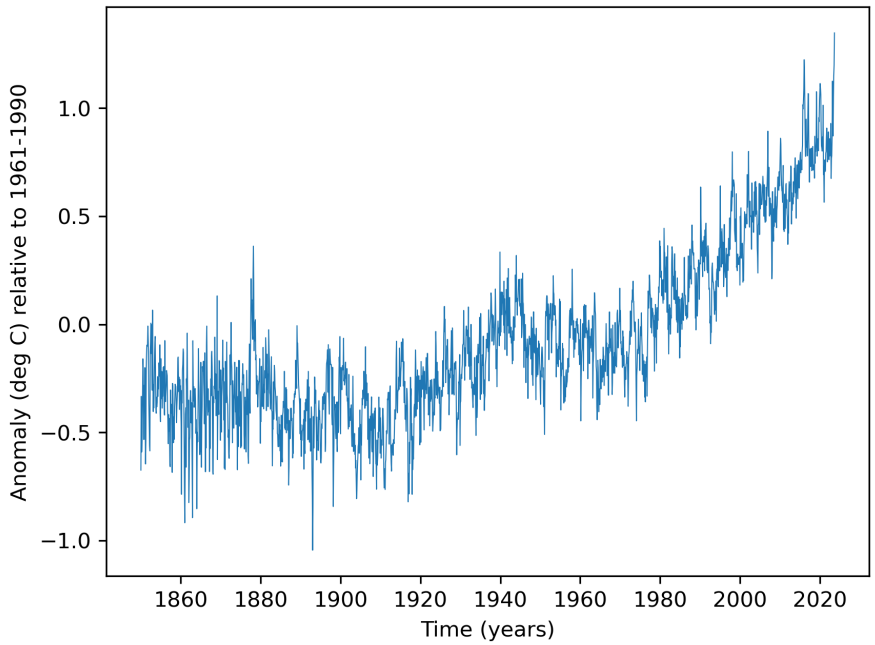

Figure 6: Caption

Articles

1. Solomon, S.; Qin, D.; Manning, M.; Averyt, K.; Marquis, M. Climate Change 2007—The Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC; Cambridge University Press: Cambridge, MA, USA, 2007;

2. Cifuentes J, Marulanda G, Bello A, Reneses J. Air Temperature Forecasting Using Machine Learning Techniques: A Review. Energies. 2020; 13(16):4215. https://doi.org/10.3390/

3. Pasini, A.; Lorè, M.; Ameli, F. Neural network modelling for the analysis of forcings/temperatures relationships at different scales in the climate system. Ecol. Model. 2006, 191, 58–67.

4. Fildes, R.; Kourentzes, N. Validation and forecasting accuracy in models of climate change. Int. J. Forecast. 2011, 27, 968–995.

5. Hippert, H.S.; Pedreira, C.E.; Souza, R.C. Combining neural networks and ARIMA models for hourly temperature forecast. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN 2000, Neural Computing: New Challenges and Perspectives for the New Millennium, Como, Italy, 27 July 2000.

6. Hossain, M.; Rekabdar, B.; Louis, S.J.; Dascalu, S. Forecasting the weather of Nevada: A deep learning approach. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015.