Searches for new physics with machine learning at the LHC

Khalil Pierre[1], Violet Hill[1], Henning Flaecher[1], Ben Krikler[1]

[1]School of Physics, University of Bristol

1. Introduction

The Standard Model (SM) is the most precisely verified theory in physics.

However, the SM is believed to be an incomplete theory. Multiple observed phenomenon cannot be explained within the SM. These include:

1. The baryon asymmetry of the universe
2. A description for gravity
3. A particle candidate for dark matter

These inconstancies lead physicist to believe that there must be physics beyond the standard model. The search for invisibly decaying Higgs bosons is an active area of research at the LHC into BSM physics.

In our project we employ machine learning as a novel technique to investigate invisible Higgs decays.

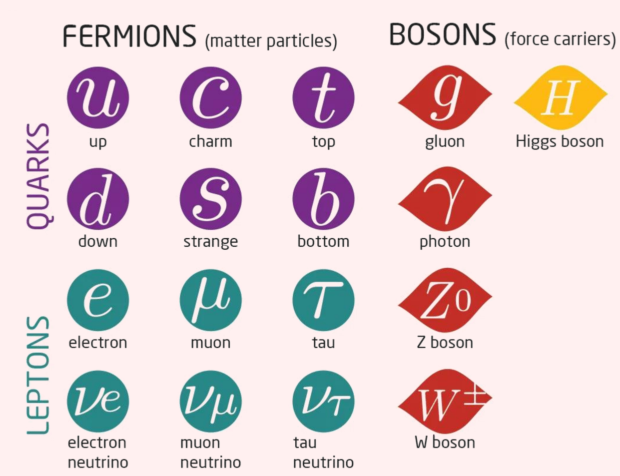

Fig 1: Elementary particles of the Standard Mode

2. Invisible Higgs

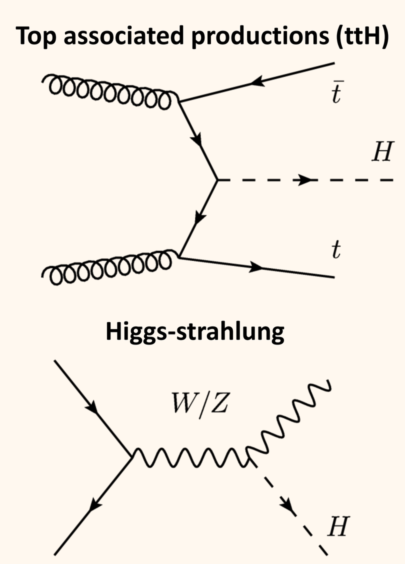

Fig 2: Leading order Feynman diagrams for the two Higgs production modes investigated.

Dark matter is observed to exist through it's gravitational effects on astronomical objects.

One of the few things astronomers know about dark matter is that it has mass. Since the Higgs boson couples to mass, massive dark-matter particles should in theory interact with it. This opens up the possibility of Higgs boson decays into dark matter particles.

Using Higgs bosons produced by the LHC, it can be investigated if there are Higgs boson decays into unknown particles invisible to the detectors. Our project centres on creating algorithms to detect these invisible decays.

Latest results from the ATLAS detector place the upper bound on the branching ratio: BR($H \rightarrow$ invisible) = 0.11 ± 0.04.

Theoretical calculations have a SM BR($H \rightarrow$ invisible) of 0.12% which leaves lots of room for possible BSM physics.

3. Detecting invisible Higgs at CMS

The compact muon solenoid (CMS) experiment is one of two large general purpose particle physics detectors at the LHC. The CMS detector has around 100 million sensors, capable of collecting terabytes of data per second. This data is used to reconstruct collision events at the LHC, and make predictions about the particle constituents of the collision. The most common object within the collisions are 'jets', which form from the hadronization of unbound colour charge. If dark matter is present in collisions at the LHC its signature will turn up in the missing transverse momentum.

To measure the invisible Higgs branching ratio, invisible Higgs decay events need to be distinguished from background SM processes with similar detector signatures. Variables like the number of jets, mass of the jets, angles between jets are used to distinguish signal from background events. Selection cuts are used to supress background events. For example, common event selection cuts for the $H \rightarrow$ invisible: ttH are as follows:

- HT > 200Gev
- Number of jets >= 5
- Number of b-jets >= 2
- MET > 200Gev
- MHT/MET < 1.25

In this research project, simulated data from the CMS detector was used. The simulation data is produced using monte Carlo methods. The probability distributions of the SM are modelled using random number generators.

In order to efficiently compute statistics on this data, each event is summarised by 12 event variables and an non-uniform number of jets.

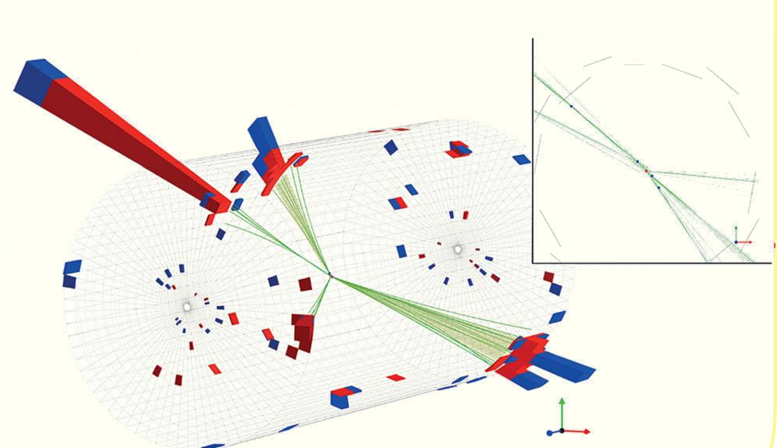

Fig 3: Jet reconstruction in simulated CMS event. Image credit: CMS Collaboration.

4. Machine learning

Machine learning allows a computer to learn the complex non-linear relations between variables in order to classify LHC events.

We employed a variety of neural network architectures as binary classifiers to distinguish *ttH → invisible* events from their associated background processes.

A dense fully-connected feedforward neural network (FFN) was used for the event level variables. Recurrent neural networks (RNN) were used for the variable length jet data.

The complete neural network architecture is shown in figure 5. This combines the FFN and RNN into a single architecture which can utilise the full dataset.

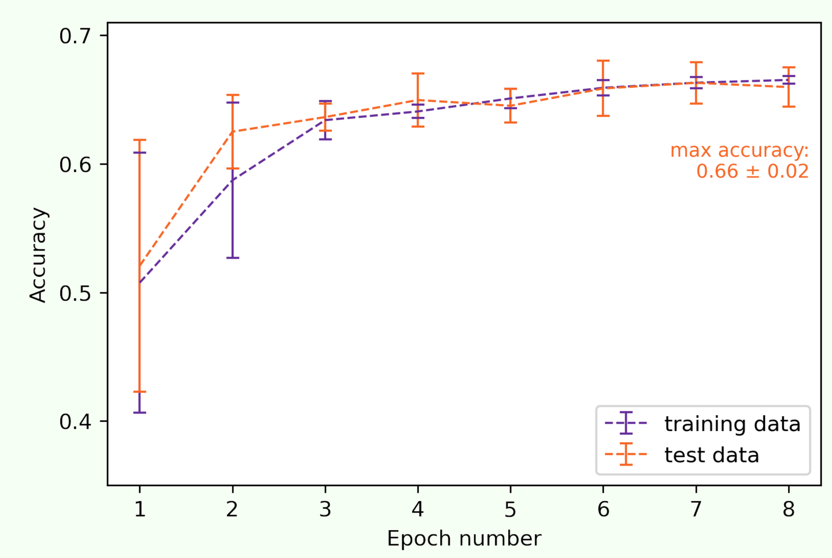

Fig 4: Training history of the RNN trained on the jet data. Average of 25 runs.

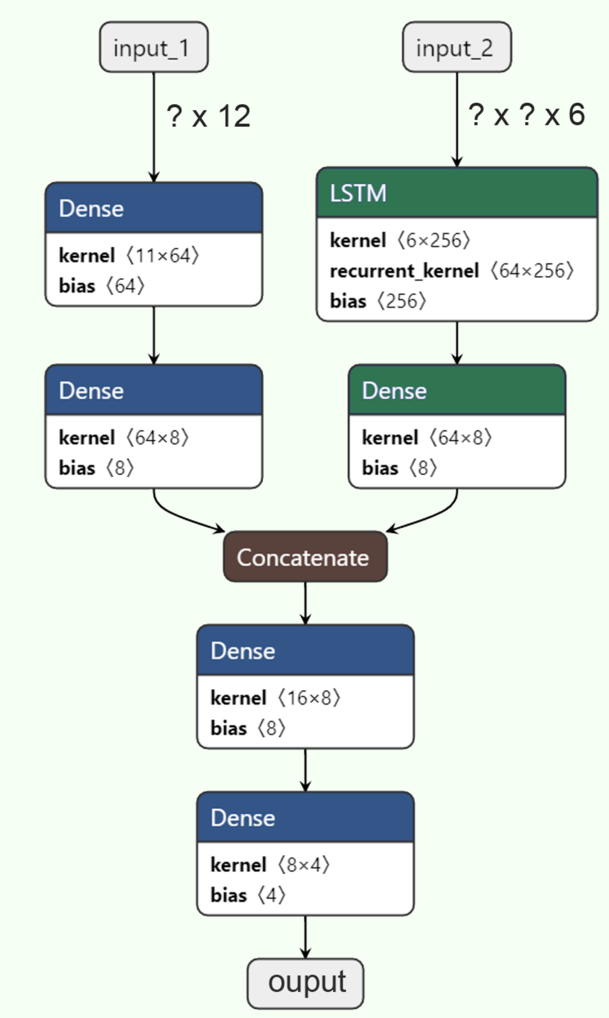

Fig 5: Combined neural network architecture.

5. Results: Feedforward network

Each model was trained on a dataset of around 400,000 simulated events.

Weights were applied to each event type to account for the relative differences in their cross-sections.

The training of each model was repeated several times, with random initialisation of the network and training data (see figure 4).

The distribution of output values of the FFN is shown in figure 6. The peaks close to 0 and 1 indicate that network is able to successfully distinguish the ttH signal from it's associated background processes.

The accuracy of a model is defined as the number of correct classifications over the total number of events. However, if a network only predicts background it can still get a high accuracy score if background events make up a majority of the events.

For a more resolved view a confusion matrix is needed. The confusion matrix is obtained by placing a threshold value on the discriminator; values above this threshold being classified as signal.

Figure 7 demonstrates how the feedforward network performed better at correctly identifying the tt-bar (background) than the ttH (signal) events.

The same plots were produced for the recurrent neural network and combined architecture. These are not presented here to avoid repetition.

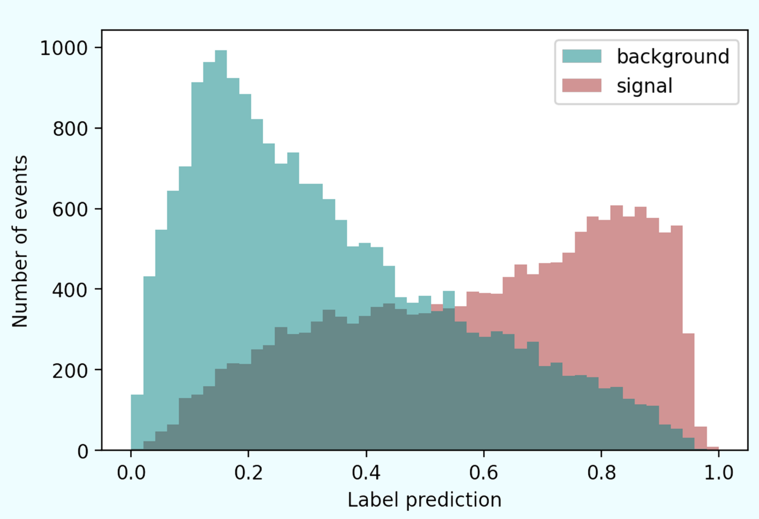

Fig 6: Distribution of labels predicted by the feedforward neural network on a test sample of ttH events.

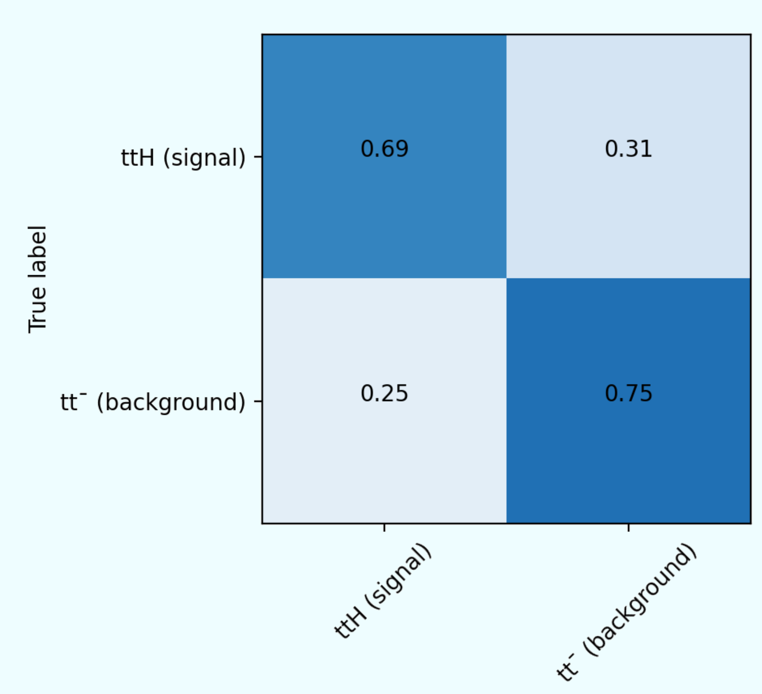

Fig 7: Confusion matrix of feedforward neural network trained on the event level variables

6. Results: All network architectures

It was observed that the RNN didn't perform as well as the FFN model. However, the combination of the FFN with the RNN produced an architecture with an accuracy higher than either network on it's own.

A peak accuracy of 0.72 ± 0.02 was obtained for the combined FFN + RNN architecture.

ROC curves are used to evaluate the ability of binary classifiers to distinguish between events (figure 9).

The area under the curve (AUC) represents the degree to which a model can separate signal from background. The higher the AUC, the greater the predicting power of the model; with a perfect network having an AUC score of 1.0.

The FFN and FFN + RNN architectures have the highest AUC scores of 0.799 ± 0.002.

Next, we added in additional Higgs boson production modes, starting with the W/Z (Higgs-strahlung) production and its associated background. This required creating multi-class classification networks.

Results for the W/Z production mode classifier are summarised in table 1.

FFN	RNN	FFN + RNN
0.91 ± 0.01	0.92 ± 0.02	0.93 ± 0.02

Table 1: AUC comparison of the neural network architecture for the W/Z events.

It was found that the ttH to invisible events were the most difficult to distinguish from background processes. The W/Z process offers a more powerful rejection of backgrounds.

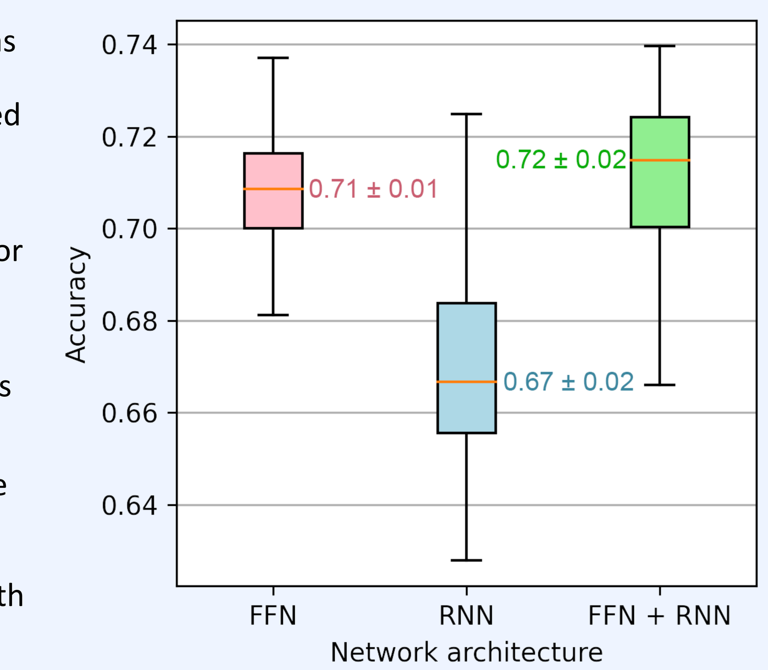

Fig 8: Accuracy comparison of neural network architectures for the ttH binary classifier.

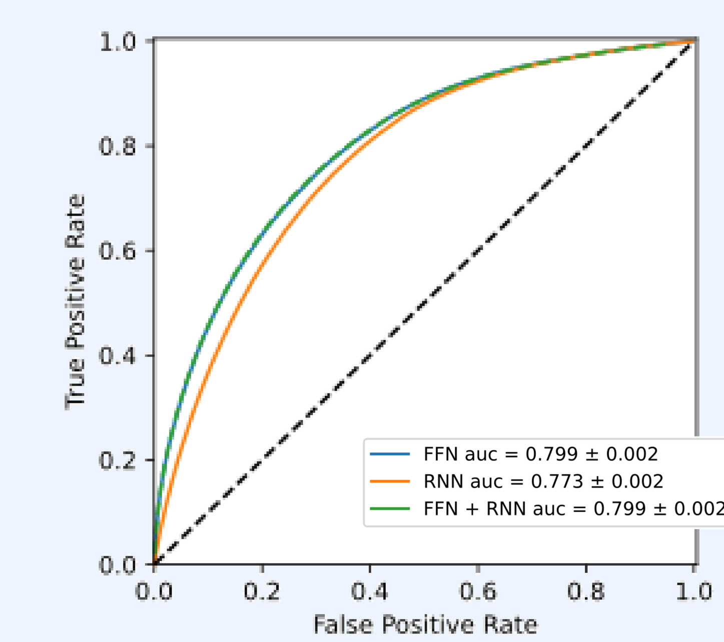

Fig 9: Average ROC curve for different network architectures for the ttH binary classifier.

7. Conclusions

We have demonstrated that machine learning is a viable tool for classifying events at the LHC.

We were able to distinguish the ttH Higgs production mode from it's associated background with an accuracy of 0.72 ± 0.02.

Feedforward neural networks and recurrent neural networks were combined into an advanced network architecture for classifying a range of data formats.

Additional Higgs production modes were included by extending the neural network from a simple binary classifier into a multi-category classifier. Improved accuracy was observed for the associated production with a W/Z boson production mode.

8. References

1. Diptimoy Ghosh et al. 'Looking for an Invisible Higgs Signal at the LHC'. In: Phys. Lett. B 725 (2013), pp. 344–351.
2. M. Tanabashi et al. 'Review of Particle Physics'. In: Phys. Rev. D 98.3 (2018), p. 030001.
3. 'Search for invisible Higgs boson decays with vector boson fusion signatures with the ATLAS detector using an integrated luminosity of 139 fb−1 '. In: (Apr. 2020).
4. 'Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment'. In: (July 2019).
5. Vishal S. Ngairangbam et al. 'Invisible Higgs search through Vector Boson Fusion: A deep learning approach'. In: Eur. Phys. J. C 80.11 (2020), p. 1055.
6. Martin Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. url: https://www.tensorflow.org/.