Poor Eyesight Chess AI Report

Alex Fish

1 Overview

Oh no, Stockfish forgot its glasses at home! The state of the art chess move-suggestion engine can only recognize the **location** and **color** of pieces on the board and must guess the piece types based on previously-observed games before it thinks of a move to make. While Stockfish with perfect vision (i.e., being told the exact board state) is far better than any human chess grandmaster, how well can it perform when it has "poor eyesight" and might incorrectly guess which pieces are on the board, influencing its decisions?

The primary goal of this project is to build a deep-learning model which uses only chess piece **location** and **color** information to predict the exact board state, trained on a history of chess games played by real people.

Even the most interesting models are mere curiosities without proper deployment, so I created a small chess program to play against this "Poor Eyesight Stockfish" locally on the command-line, instructions for which can be found on the GitHub page. Additionally, I created a Lichess "bot" account, PoorEyesightBot, where you can play against Poor Eyesight Stockfish online with no downloads or setup required! An account is required to challenge the bot and it accepts challenges with between 1 and 15 minute base time and between 0 and 10 second turn-increment time.

In Section 2, I will provide a brief overview of the relevant terminology. In Section 3, I will describe the model motivation and structure in more depth. In Section 4, I will discuss the training and testing data source, mining, storage, and volume along with a brief analysis of the data. In Section 5, I will go into the technical details of the model training, testing, and tuning process. In Section 6, I will discuss implementation considerations for using this model together with Stockfish. Finally, in Section 7, I will assess model performance and give conclusions.

1.1 Keywords

data mining, feature engineering, classification, class imbalance, deep learning, parameter tuning, ensemble modeling, model deployment.

1.2 Libraries Used

pandas, python-chess, Matplotlib, PyTorch, NumPy

2 Chess Terminology

In this Section, I'll define the relevant terminology. First, the relevant chess terminology to building the model.

Term	Description
Piece	Both players in a chess game start with the following pieces: 8 pawns, 2 knights, 2 bishops, 2 rooks, 1 queen, 1 king.
Square/Piece location	A chess piece is located on a square of the chess board. There are 64 squares in an 8x8 grid, with columns or "files" labeled "a" through "h" (left to right), and rows or "ranks" labeled "1" through "8" (bottom to top). For example, the white queen starts on "d1."
Piece color	One player controls the white pieces, while the other controls the black pieces. The white pieces start on the first and second ranks, while the black pieces start on the seventh and eighth ranks.
Piece type	For example, "pawn," "knight," or "rook."
Board state	The complete set of information of piece locations, colors, and types across the whole board. When any of this information changes (such as after a piece moves), it is considered a different board state.
Square occupant	The piece that resides in a square, or "empty." For example "white rook" or "black pawn."

Next, some auxiliary terminology.

Term	Description
Stockfish	A chess AI which is the state of the art in chess move decision-making.
Poor eyesight	Limiting the information available to only the piece location and color, simulating blurry vision in real life.
Perfect vision	No limitations on board state information.
Pawn promotion	If a pawn moves 7 squares forward, reaching the end of the board, it promotes to another piece. This piece can be a knight, bishop, rook, or queen and the choice is up to the player who controls the pawn.

3 Model Motivation and Structure

3.1 Motivation

As mentioned in Section 1, the goal is to build a model which can only observe chess piece location and color information to predict the full board state (i.e., predicting which exact piece is on each square).

This project is inspired by one of the algorithms in the YouTube video 30 Weird Chess Algorithms by Thomas Murphy VII, also known as "tom7". However, his model only used piece location information without considering piece color, which struck me as unmotivated and arbitrary. Tom's model, as mentioned in the video, has deep faults in that it tends to think any piece deep into a given side of the board belongs to that side's starting color, allowing his queen to infiltrate the AI's side of the board without any resistance.

3.2 Structure

The modeling is actually done through an ensemble of 64 smaller classifiers, one dedicated to each square on the board. Each of these 64 models uses the entire board state to predict which piece is on its dedicated square (with a "no piece" designation in case there is no piece on the square).

These models have 128 binary inputs: a binary input for whether a white piece is on a given square and a binary input for whether a black piece is on a given square, for each square on the chess board. When the given square is empty, both inputs for the square are set to zero.

One sample of input data to the model (representing the poor eyesight board state information) for an unmoved board has the following structure.

feature	value
a1_white	1
a1_black	0
a2_white	1
a2_black	0
...	...
e4_white	0
e4_black	0
...	...
h7_white	0
h7_black	1
h8_white	0
h8_black	1

The models have either 11 or 13 possible outputs. A given square can contain a white or black pawn, knight, bishop, rook, queen, king, or can be empty. However, the squares on the first and eighth ranks cannot ever contain a pawn because pawns start on the second and seventh ranks, only move toward the other side of the board, and promote to another piece when they reach the opposite end of the board.

One sample of output data (representing the correct piece) for the "a1" square on an unmoved board has the following structure, note that there are 11 total possible occupants of the square, due to the impossibility of pawns.

output class	value
a1_white_knight	0
a1_black_knight	0
a1_white_bishop	0
a1_black_bishop	0
a1_white_rook	1
a1_black_rook	0
a1_white_queen	0
a1_black_queen	0
a1_white_king	0
a1_black_king	0
a1_empty	0

One sample of output data (representing the correct piece) for the "e4" square on an unmoved board has the following structure. Note that there are 13 total possible occupants of the square, due to the possibility of pawns.

output class	value
e4_white_pawn	0
e4_black_pawn	0
e4_white_knight	0
e4_black_knight	0
e4_white_bishop	0
e4_black_bishop	0
e4_white_rook	0
e4_black_rook	0
e4_white_queen	0
e4_black_queen	0
e4_white_king	0
e4_black_king	0
e4_empty	1

The hope is that these models they will learn common board patterns and structures. For example, if only a few pieces have been moved from the second rank, then those pieces are likely pawns, if there is a piece far beyond the back rank while the second-from-back rank is still occupied and not many pieces have moved, then the further piece is probably a knight or bishop, or if there are diagonally-adjacent chains of pieces, then they are probably pawns, and so on.

One fault of this model is that the squares do not work together to form a sensical, legal board state. The individual square models predict their occupant independently. It is possible no king is predicted, multiple kings could be predicted, a large number of queens could be predicted, etc. The predicted board state is minimally algorithmically fixed to produce a legal board state for playing against Stockfish.

The better this model performs, the better Stockfish's idea of the exact board state, the tougher it will be to beat!

Because there are an incredible amount of possible board states (the best estimates place this number around 10^{45}), the models cannot possibly be trained on all of them. I focus the training on positions played in actual games. These will be more useful when using the model to play against humans as it has observed relevant positions.

3.3 Why not predict a whole boardstate at once?

It is possible to predict an entire board state at once with careful setup of the model's cost function, using only segments of the then-800-length output layer (considering 11 possible occupants for the 16 squares on the first and last ranks, and 13 possible occupants for the rest of the 48 squares) to predict a given square's occupant. However, the network would have to be quite large to accommodate for the vastly increased output layer size. Also, as will be discussed in Section 5.2, each square is almost always empty so dealing with the large "class imbalance" square-by-square would be a huge challenge when predicting an entire board at once.

3.4 Incorporating the Model with Stockfish

Figure 1 shows the workflow chart of a game when Poor Eyesight Stockfish plays against a game against an opponent. The opponent makes a move, then the poor eyesight board state (piece location and color data) is sent to the poor eyesight model. The poor eyesight model predicts a board state, then this predicted board state is fixed as discussed in Section 6, ensuring one king of each color is on board. Stockfish uses this fixed predicted board state to make a move. This move is made, and it is the opponents turn again. This repeats until the end of the game.

4 Data

4.1 Data Source and Parsing

The training and testing data come from Lichess's database of rated games. I semi-arbitrarily chose the March, 2016 database. This database contains PGN data for 5,801,234 games. When unzipped, this database file is 5.07GB. A PGN (Portable Game Notation) file contains information on the chess game played, including player usernames, player ratings, the game result, the moves played, etc. I used the library python-chess to parse these files and analyze individual board states by progressing a chess board through a each game's list of moves.

4.2 Data Storage and Volume

For each game in the dataset, I parsed board state information from every turn. After every turn, I stored the piece and color on every square (with an empty designation) and stored the result as one row in a table.

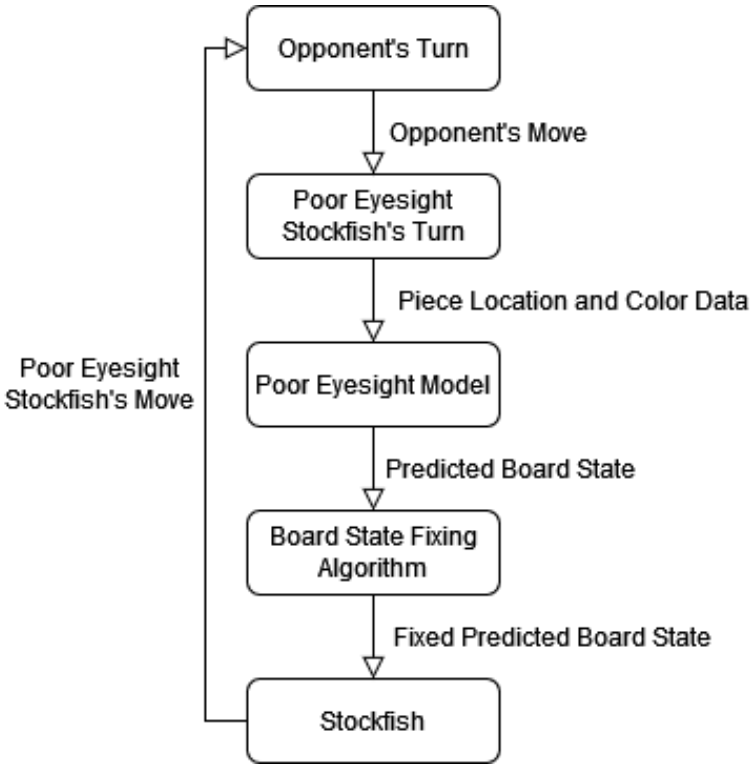


Figure 1: Chess game workflow chart with Poor Eyesight Stockfish

Because, especially in the first few moves of the game, there are very common board states, I only stored unique board states. It did not make sense to store an unmoved board or one with a single pawn moved thousands of times.

Additionally, after the first 10 moves or so, it is very *unlikely* a board state had been seen before. Thus, a huge number of board states were observed. Because I have limited storage and RAM on my computer, I had to limit the total board states stored in some structured way. I decided to only look at 100,000 total games in which both players were rated at least 2000. The board states will be observed from skilled players and thus be more likely or more at least important to get right. I figured observing lower-rated players' games would be more likely to lead to fairly nonsensical and unlikely board states.

This produced just over 6.5 million unique board states. Loading all of these into a pandas dataframe at once used approximately 20GB of my 32GB total RAM. When simply loading the dataset takes so much RAM, it is difficult to fit in model training. I trimmed off a minor amount of board states from the end to bring the dataset to 6.5 million unique board states, randomly shuffled them, then stored them in 65 100,000-board state files. Splitting the files made it easier to manage analysis and testing than loading one large file in chunks

every time.

If I had access to larger computing nodes, more RAM, more storage, more time, etc. combined with a more suitable distributed database library such as Apache Spark, I would feel more comfortable training on more board states.

4.3 Data Analysis

The data collected contains 6.5 million unique board states from 100,000 chess games played by users of Lichess rated at least 2000. Figure 2 shows the distribution of total pieces on the board in these games. There's a minimum of two pieces (the two kings), a maximum of 32 (a board with no pieces captured), and a mean of 21.1 pieces. There's a distinct pattern of even-valued total pieces being far more common than odd-values total pieces, at least when total pieces is greater than 20 or so. This is likely due to piece trading (capturing an opponent's piece with one piece, then that piece immediately being captured in return) being more common than winning a piece (capturing an opponent's piece without them able to recapture), at least early enough in the game when there are still many pieces remaining on board.

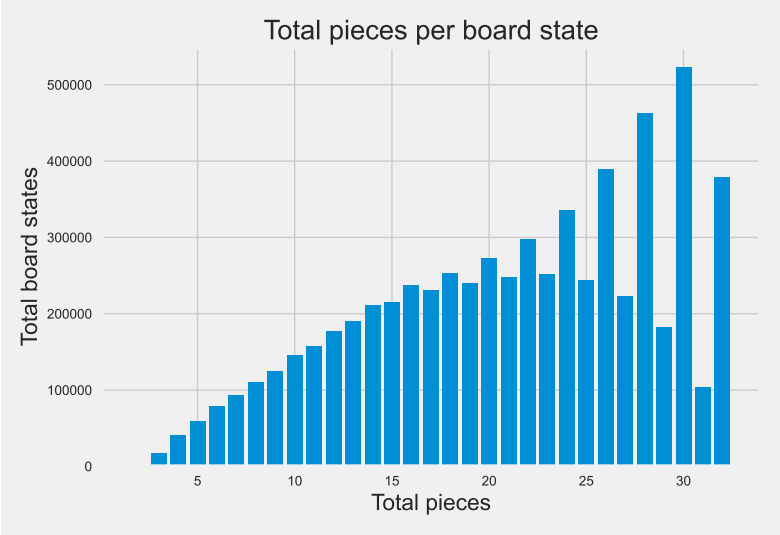


Figure 2

There are too many squares to draw any meaningful analysis if looking at all of them. Narrowing our scope to a few squares scattered across the board, "a1," "e1,", "e4," "d5," "h5," and "g8." Figures 3-8 show the distribution of pieces on these squares. Note that the y-scale is logarithmic because the squares are empty or sometimes contain a particular piece the vast majority of the time.

All of the squares are most commonly empty. "a1" has a white rook fairly frequently, which is expected because the rook starts on that square and will

rarely move until castling kingside, which might not happen. "e1" commonly has a white king and white rook, which is expected because the white king starts on "e1" and the white rook is commonly moved from "f1" to "g1" after castling, where the rook shines as the "e" file is commonly unblocked by pawns. Both of these squares very rarely have a black king, because it would be dangerous and unlikely for the black king to make it all the way there outside of a late endgame "fighting to promote a pawn" situation. "a1" also rarely has a knight because there are almost always more useful squares for knights to be. Neither of these squares ever have pawns for reasons discussed above.

"e4" and "d5" both commonly have pawns, with the more common color being the color which starts closer to them. Pawns are used to fight over the center early and may stay there a while, or at least have pawns from nearby files recapture to maintain a pawn on those files. Bishops, knights, and queens commonly pass through the center squares.

"h5," on the edge of the board, is most commonly empty. Controlling the center is considered important, so maintaining pieces on the edge would be a waste of time. Some pieces, such as knights and bishops, will pivot on these squares to get to more useful squares. Pawns are the most common pieces on this square, as they are used in attacking a castled king or pushed to stop an opponent's pawn advance.

Black kings and black knights are the most common pieces on "g8." This is the square where a black knight starts, and the square where the black king lands when castling kingside, which is the most common form of castling. Similar to "a1," the white king is very rarely on this square and is likely only there in a late game situation.

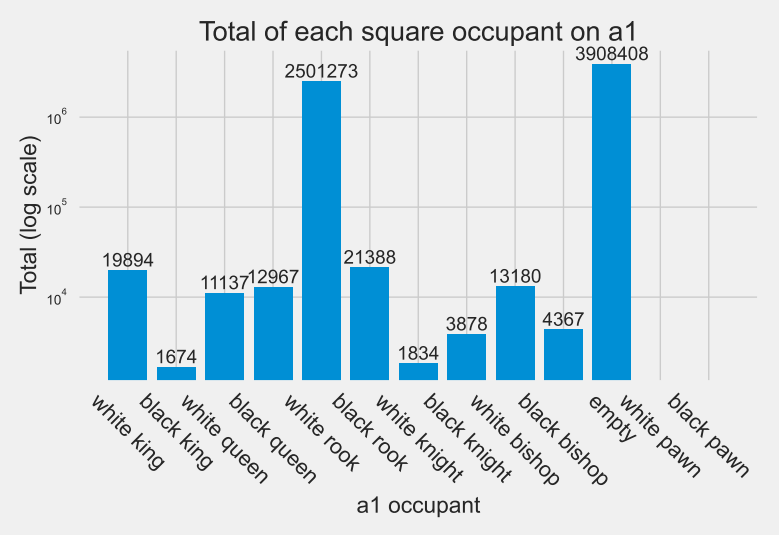


Figure 3

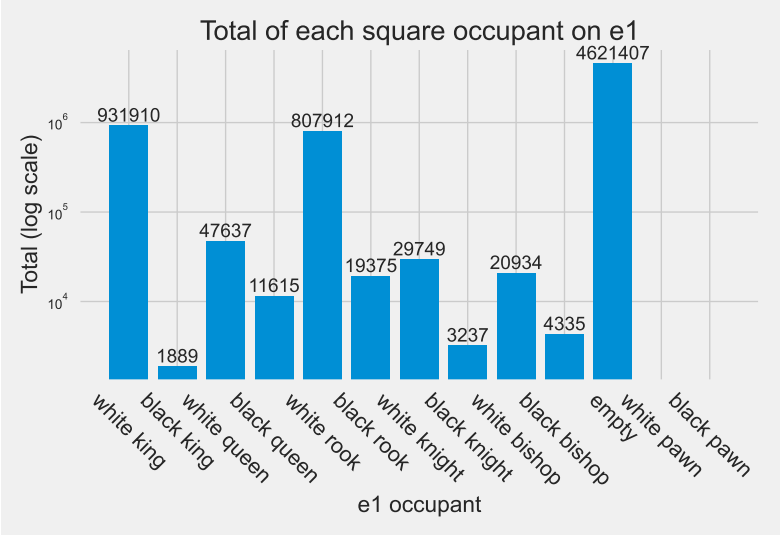


Figure 4

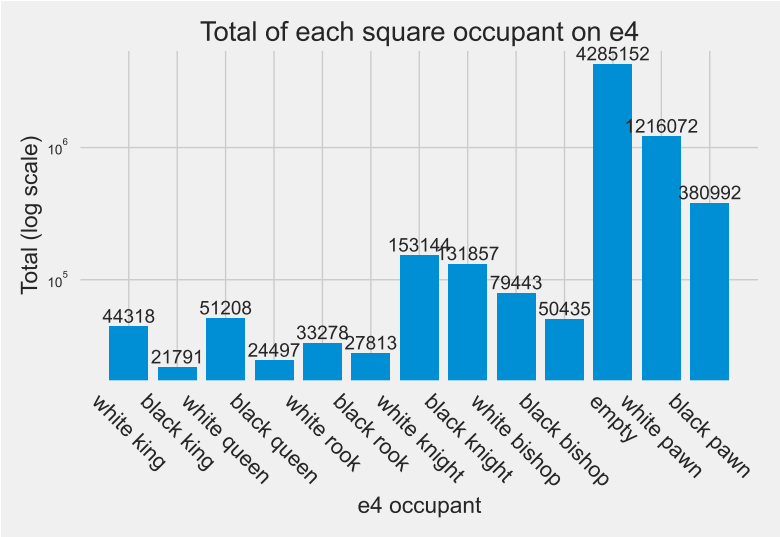


Figure 5

5 Training, Testing, and Tuning

I used 5.5 million unique board states for training and the remaining 1 million unique board states for testing.

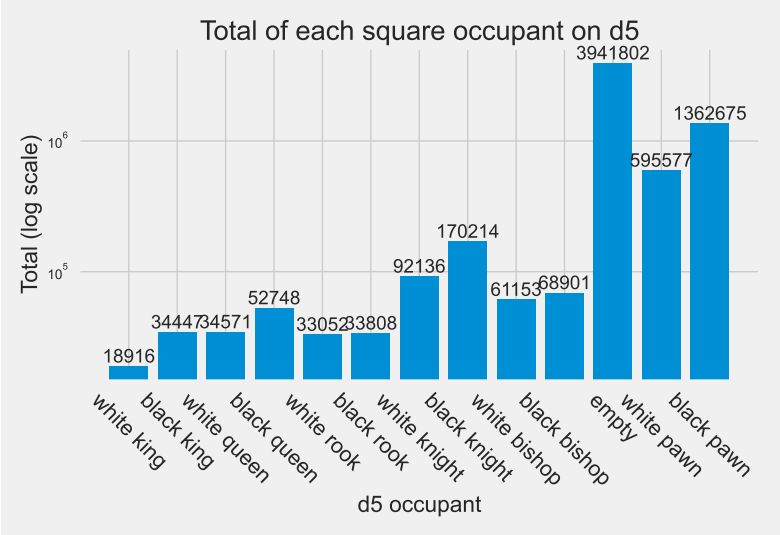


Figure 6

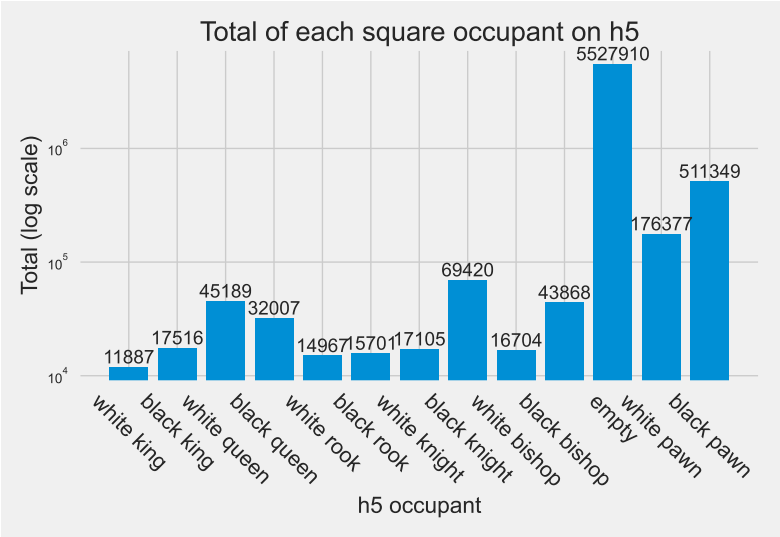


Figure 7

5.1 Parameter Tuning

I used deep learning models, built with PyTorch, for each square's classification model. Due to lengthy training times bogging down my personal computer, I decided to tune parameters manually. I examined performance on a few squares, "f5" and "g4", squares on which a variety of pieces commonly land. If I had

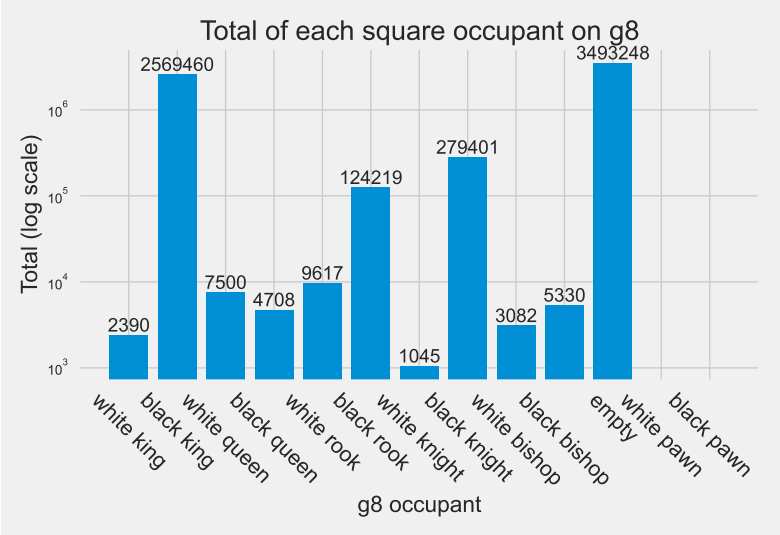


Figure 8

access to larger computing power, I would use a hyperparameter tuning library such as Optuna to find the most optimal parameters.

I started with a similar model structure to that of tom7 in 30 Weird Chess Algorithms: a three layer neural network with a moderately large first layer, a much larger second layer, then a third layer which was equal in size to or slightly smaller than the second layer. I kept all layer sizes equal to a power of two, with all three layers containing at least 128 nodes, the size of the input layer. I examined a variety of structures, with the first layer ranging from 256 to 2048 nodes, the middle layer ranging from 512 to 8192 nodes, and the last layer ranging from 128 to 2048 nodes. I found that the performance of the (512,2048,2048) had the best performance. All of these evaluations were performed using a batch size of 1000 and 5 training epochs. A learning rate of 0.1 provided the fastest convergence without any divergence issues after testing with 0.1, 0.05, 0.01, 0.05, 0.001. Using a step size momentum of 0.9 sped up convergence drastically.

Because I expected the last layer to be significantly smaller than the second layer, somewhere between the prior layer and the output layer, I decided to add fourth layer. I kept the first 3 layer sizes and evaluated with a fourth layer sizes of 128 through 1024. A layer size of 256 provided the best performance.

Finally, I evaluated performance on the test set across a number of total training epochs to avoid over-fitting. I found that 20 epochs had good performance, with more epochs leading to overfitting issues and decreased test set accuracy.

Figure 9 shows the final neural network architecture for a single square's model.

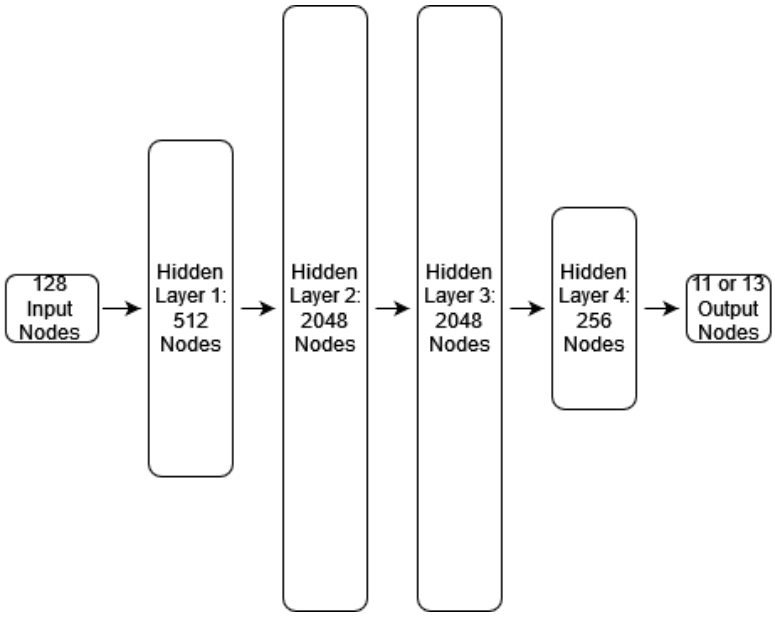


Figure 9: Poor Eyesight neural network model for predicting the occupant of a single square.

5.2 Class Imbalance

For any given square, there is no piece on the square for between 60-95% of the board states. This provides a challenge for training, as the models tended to predict "empty" every time and achieve low training cost. To solve this, I over-sampled the board states, split by the corresponding square occupant, by repeatedly training on them in inverse proportion to how frequent they were. The number of times to repeat training on board states corresponding to a given output class was given by 40,000 divided by the number of board states associated with given output class, rounded up. This process of repeating samples was performed for each of the files originally containing 100,000 board states used in training.

This led to an overall minimum of approximately 2.6 million samples of each square occupant, and an overall approximately five times increase in the number of samples trained per epoch, for a total of 31 million samples per epoch. This roughly balances the number of dataset entries of each square occupant such that each square occupant comprises approximately 1/13 of the dataset, where 13 is the typical number of possible square occupants.

This over-sampling process during training drastically improved test set accuracy, going from predicting "empty" every time to a per-piece accuracy over 60%, sometimes over 90%. Note that the test set was *not* over-sampled.

There are methods of creating interpolated data to oversample and reduce

overfitting and often involve some sort of K-means clustering. Because of a variety of factors—binary input data, desire to include only board states seen in real games, the finicky nature of having 13 output classes, the drastic nature of the class imbalance—I decided to do a simple repeated-sample technique. The performance increase on the test set speaks to the success of this technique in this case.

6 Implementation Considerations

As mentioned in Section 3, the individual models predicting the piece on each square do not work together to ensure a legal board state. When assessing accuracy and performance of the individual models, this is not important. This becomes important when feeding a predicted board state into Stockfish, which requires a legal board state to suggest a next move. Fixes need to be manually/algorithmically implemented to issues preventing illegal board states.

The primary issue to fix is the amount of kings on the board. For example, if there are no kings predicted, the square which is *most likely* a white king is set to contain a white king. If there are multiple white kings predicted, the square which is most likely a white king is kept, while the less likely squares are set to the next-most likely piece. The same process is performed for black kings.

If Stockfish suggests a move that is illegal given the true board state, a random move is made for it.

Due to poor performance of Poor Eyesight Stockfish, I also implemented a "memory." This memory consists of the last 3 moves Poor Eyesight Stockfish made would be "remembered" and modified in the board state's output if the pieces moved had not been captured and the models incorrectly predicted the recently-moved pieces. After all, people with very poor eyesight would still remember the last few moves they made. I did not want to include *too much* memory, as that would defeat the purpose of this project.

The memory did not help very much and will be discussed more in Section 7.2.

7 Performance

7.1 Model Performance on Testing Data

I tested the models on 1 million board states which they had not been trained on. The models did very well on the test data given the drastic class imbalance. Figure 10 shows the frequency of the number of incorrectly-predicted squares per board. 21% of the time, the models get the entire board exactly correct, while 70% of the time there are 2 or fewer incorrect square occupants predicted and 86% of the time, 3 or fewer.

Figure 11 shows the board same board state accuracy, but looking at a subset of the board states. Figure 11a shows the accuracy on boards with more than 24 pieces on the board (at least 75% of the original total pieces), and Figure

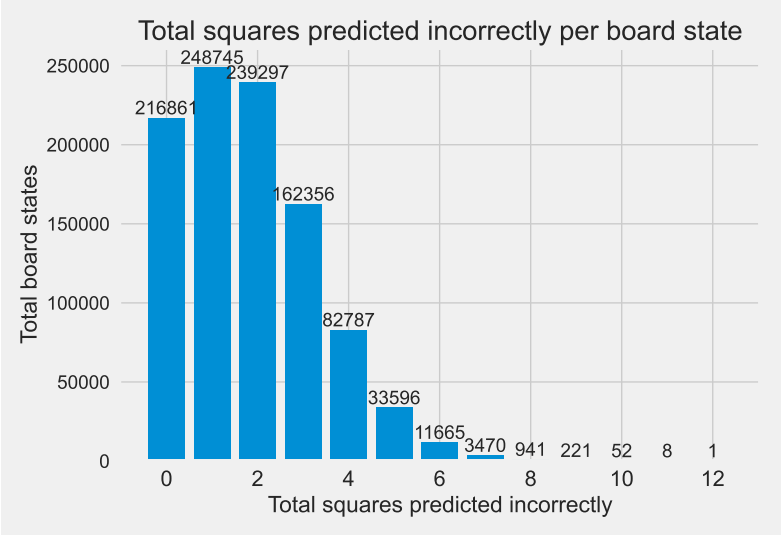
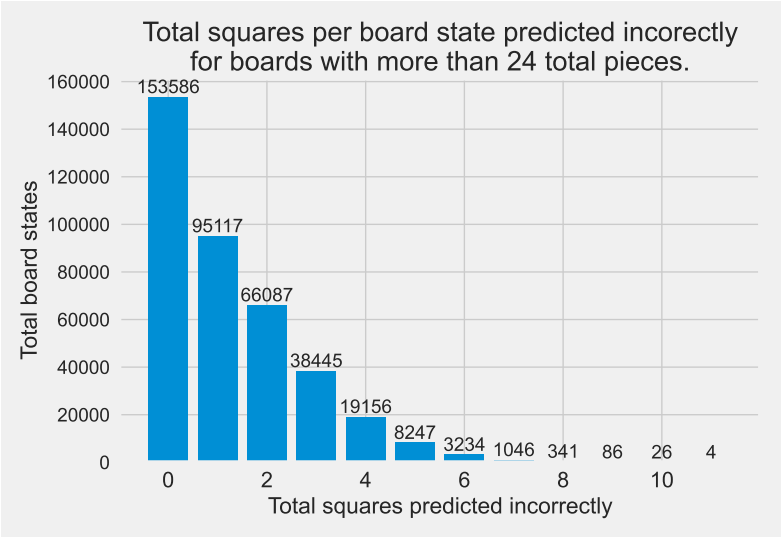


Figure 10

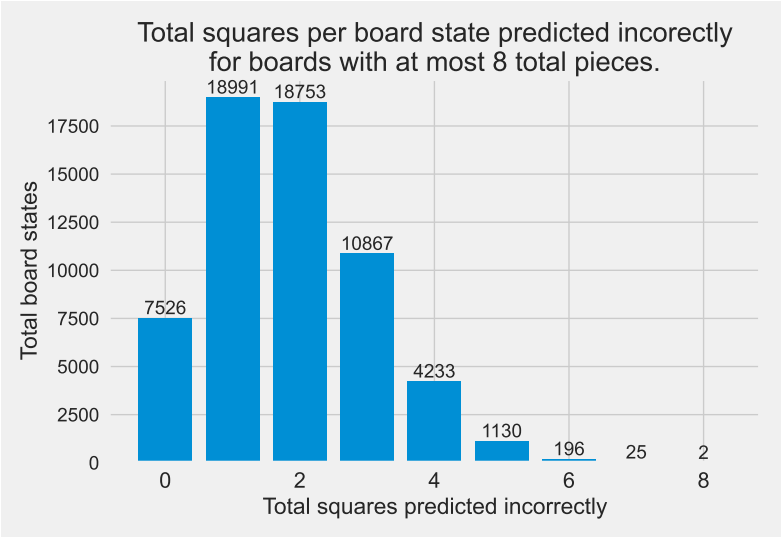
11b shows the accuracy on boards with 8 or fewer pieces on the board (at most 25% of the original total pieces). The models are more accurate than average when there are more than 24 pieces on the board, and less accurate when there are 8 or fewer pieces. If there are still a lot of pieces on the board, they are likely in expected positions relative to other pieces, while if there are very few pieces on board, they have much more freedom of movement and there can be a large variety of remaining pieces. It would be very hard to tell if there are 4 pieces on each side if a piece is a pawn, bishop, knight, etc. Additionally, it is harder to get a lot of training data for board states with few pieces because of the huge number of reasonable piece arrangements on the board. The large number of possible reasonable board states combined with the variety of pieces which could remain makes modeling the endgame a huge challenge. This comes into play when actually using Poor Eyesight Stockfish, discussed in Section 7.2.

Figure 12 shows the mean pieces predicted incorrectly versus the total non-empty squares on the board. The mean is roughly equal to 2 for any amount of non-empty squares. The mean decays as the number of pieces approaches 32 (zero pieces captured) and 2 (only the kings remaining on board). The mean is nonzero for two non-empty squares clearly showing the model's fault in not always producing a legal, sensical board state in that there is only one combination of pieces when two pieces are on board (the white king and black king in a stalemate). The standard deviations of the incorrectly predicted squares are between 1 and 1.5 for all but 2, 3, 4, 31, 32 non-empty squares.

The mean is fairly large for three to five non-empty squares, showing both the difficulty in predicting the last few remaining pieces and the need for more training data in this endgame situations. When there are only a few pieces


Total squares per board state predicted incorectly
for boards with more than 24 total pieces.
Total board states
160000
140000
120000
100000
80000
60000
40000
20000
0
153586
95117
66087
38445
19156
8247
3234
1046
341
86
26
4
0
2
4
6
8
10
Total squares predicted incorrectly

(a)


Total squares per board state predicted incorectly
for boards with at most 8 total pieces.
Total board states
17500
15000
12500
10000
7500
5000
2500
0
7526
18991
18753
10867
4233
1130
196
25
2
0
2
4
6
8
Total squares predicted incorrectly

(b)

Figure 11

left on the board, there are so many possible sets of pieces which a player could use to achieve a checkmate while the losing player could have practically any leftover pieces as they get checkmated. Also, there are comparatively few training examples from the endgame because the game could end before such a late situation and the game usually ends quickly once it gets to that point.

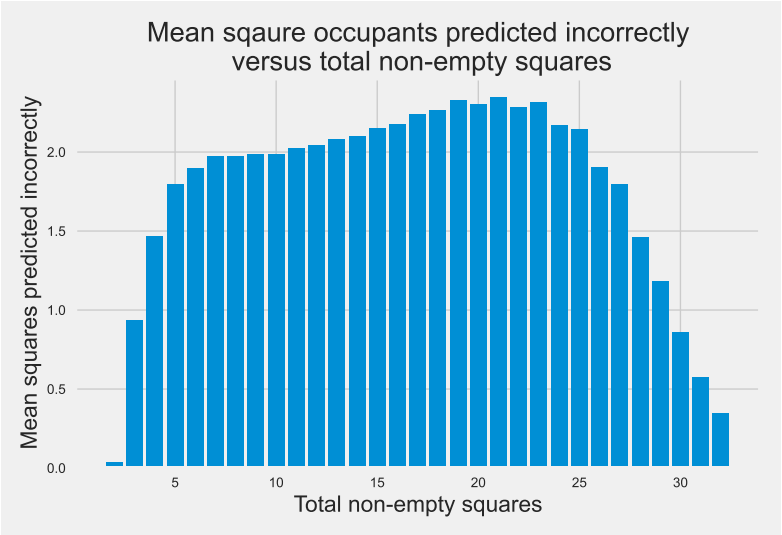


Figure 12

Figures 13-18 show the per-piece accuracy of the same squares analyzed in Section 4.3, "a1," "e1,", "e4," "d5," "h5," and "g8." Some squares fail to ever predict a piece correctly, such as the failure to ever correctly predict a black queen on "e1". The accuracies are very high for the common pieces and lower for the less frequent pieces, likely indicating a deficiency of the repeated-sample oversampling technique in training. The models were only used to seeing certain pieces on their square in particular board states. There is also a factor of seeing another piece more commonly in similar board state setups, leading to unavoidable errors.

The fact that the models have such high accuracy on a variety of pieces– the "e4" model has over 70% accuracy on 5 pieces, including pieces of both colors–shows some form of success. Even though the "d5" model failed to ever accurately predict white rooks or black knights, it had over 70% accuracy on 6 pieces. The "g8" model correctly predicted pieces at least 55% of the time, with 3 pieces having at least 95% accuracy.

Table 1 shows the overall proportions of predicting a color piece (or empty), when the actual piece is a given color (or empty) across all squares and board states. Because the models receive piece color information on their corresponding square (along with all other squares), we would expect this to be perfect or nearly perfect. This is confirmed.

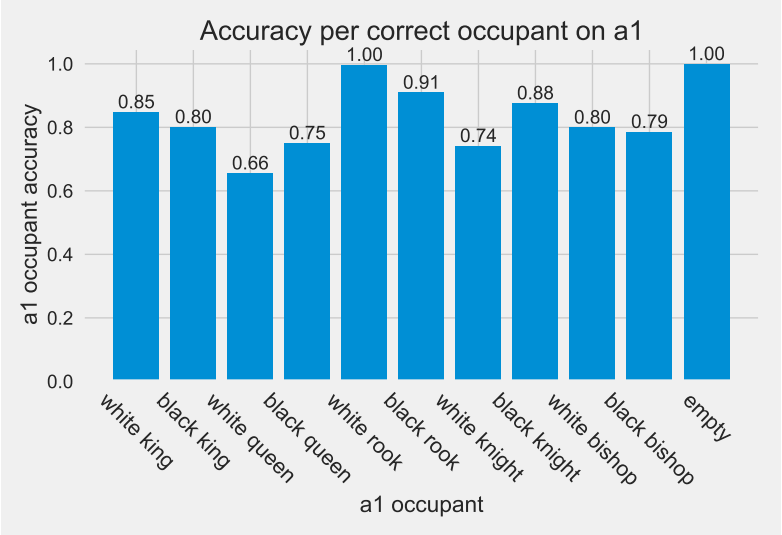


Figure 13

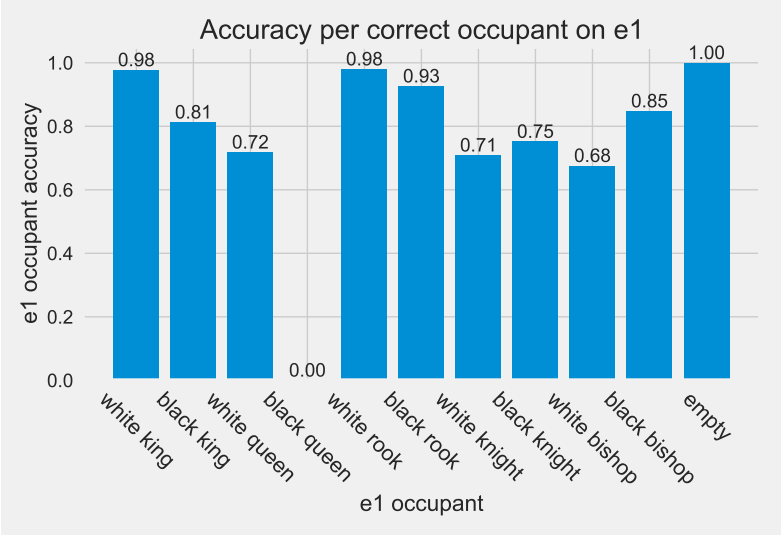


Figure 14

7.2 Poor Eyesight Stockfish Performance

Poor Eyesight Stockfish performs quite badly in real games. It is able to play perfectly or nearly perfectly for the first 10-15 moves because the models are able to predict the board with sufficient accuracy for Stockfish to be able to recommend a good move. However, eventually, as the board state gets more

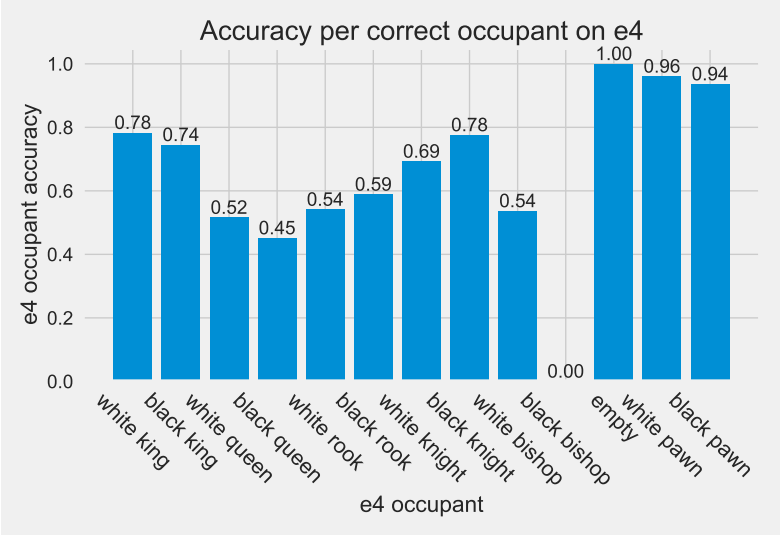


Figure 15

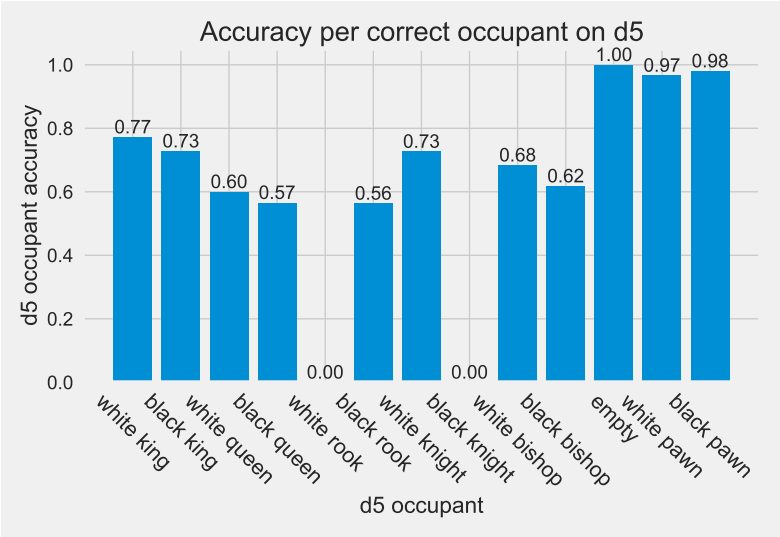


Figure 16

complex, the models start to lose their accuracy. At this point, Stockfish has starts to have such a misunderstanding of the true board state, it suggests moves that are illegal, and a random move is made for it. Once it starts playing random moves, practically any player or other bot account could beat it.

Performance did not improve with the "memory" feature addition, mentioned in Section 6. It played around 15 games against progressively weaker

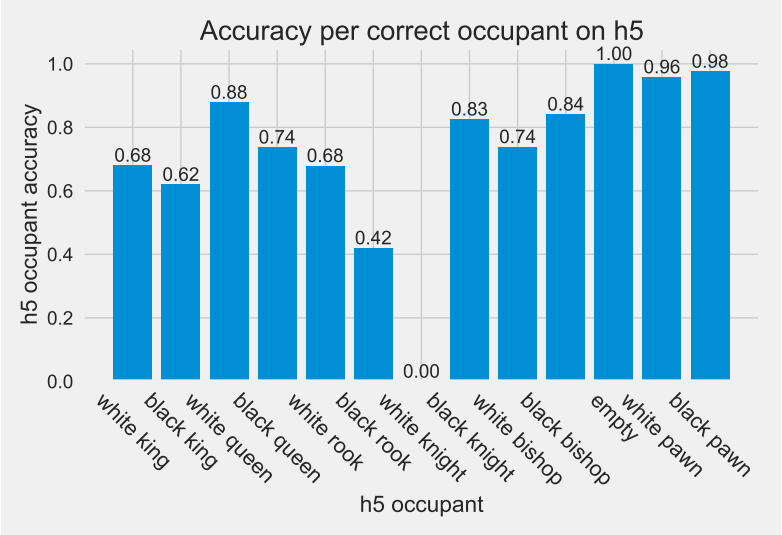


Figure 17

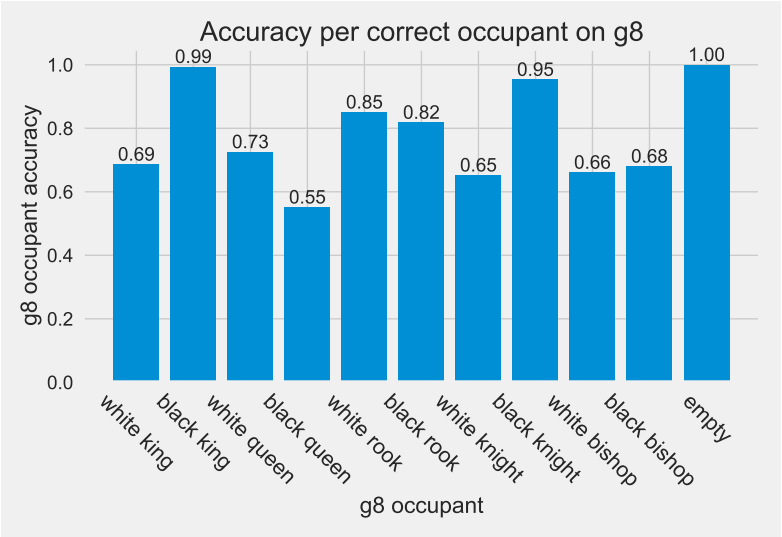


Figure 18

opponents (due to its own rating dropping) with the memory feature and did not win once. So far, the only win attributed to Poor Eyesight Stockfish on Lichess is when I ran out of time while play-testing and fixing bugs!

Figure 19 shows the typical story of a game played by Poor Eyesight Stockfish. This plot shows a post-game computer evaluation move-by-move of the game. A more positive evaluation indicates white is winning, while a more neg-

Predicted \ Actual	White	Black	Empty
White	0.999351	0.000431	0.000218
Black	0.000447	0.999278	0.000265
Empty	0.000033	0.0000025	0.999994

Table 1: Ratio of colors (or empty) predicted for each actual color (or empty) across all 1 million testing board states and all squares.

ative evaluation indicates black is winning. Early in the game, the evaluation climbs in Poor Eyesight Stockfish's favor, as the models produce an accurate enough board state for it to act close enough to regular Stockfish, which is far stronger than the opponent "maia5," a bot trained to play like intermediate human players.

The first major dip, at move 14, happens when Poor Eyesight Stockfish does not move its bishop from being attacked by its opponent's pawn. The evaluation climbs back a little when maia5 chooses to take a pawn with its knight, positioning itself for an attack on Poor Eyesight Stockfish's queen, when taking the bishop would've been better. However, Poor Eyesight Stockfish does not realize this black piece is a knight, so doesn't move its queen. It chooses to move a pawn to attack maia5's knight instead, leading to the second major dip in evaluation. maia5 takes the queen. An unhindered Stockfish may be able win against maia5 from this position, but at this point Poor Eyesight Stockfish does not stand a chance. The record of this game can be found at https://lichess.org/P0DUK5dm#1, where you can use the left and right keyboard arrowkeys to go through the game's moves.

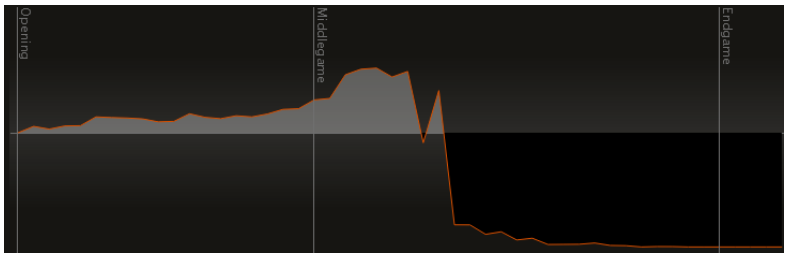


Figure 19: Post-game computer evaluation of a game played between Poor Eyesight Stockfish (white) and the bot "maia5" (black). A more positive evaluation indicates white is winning, while a more negative evaluation indicates black is winning.

7.3 Performance Conclusions

The models did quite a good job of predicting the board state in general. However, there is clearly a deficiency as evidenced by the poor performance of Poor

Eyesight Stockfish. It performs excellently for the first 15 or so moves, which is 30 different board states. However, making one mistake (such as not realizing its queen is under attack) is enough to snowball each game into an inevitable loss.

This could be partly solved by training on more board states, for more time, with a larger network, with a more thorough parameter tuning process. In the end it will never be perfect, for a few reasons. First and foremost, it is attempting to predict the full board state with incomplete information. It is bound to get information wrong. Secondly, there are too many possible board states to ever train on all of them, so there are bound to be blind spots. Finally, especially in the middle and late game, after each player has moved all their pieces, pieces have been taken, and attacks have been formed, there is too much flexibility to get accurate predictions all the time. A bishop and knight could swap places, or an unexpected combination of pieces could be left on the board near the end of the game. We can see this flexibility causing issues in the models–the models on the edge of the board ("a1," "e1", "g8") have much better performance than those at the center ("e4," "d5")

This was a fun project to see how well a model could predict the exact state of "reality" with deliberately incomplete information. I also had a lot of fun setting up the Lichess bot and watching my own creation automatically play games while I monitored it on the backend. I was pleasantly surprised at the performance it did achieve, and enjoyed seeing where and how it *failed* despite successes in other cases.