Project Title: NBA Game Match-up Prediction Tool
Project Group #: 52
Project Team Members: Michael Lamontagne (mlamontagne6): (M.L.), Ethan Garbow (egarbow3) (E.G.), Diego Escalera (dayala9) (D.E.), Sebastián de la Hoz (sluna7) (S.H.), Dan Goldin (dgoldin3) (D.G.), Yotan Demi-ejegi (ydemiejegi3) (Y.D.)

Introduction & Project Motivation

According to ESPN, NBA Teams and the sports betting industry spend well in excess of $3.1 billion in research surrounding player stats and predictions in order to gain a competitive advantage in the respective fields. This topic is of high interest and our objective is to provide a device for fellow NBA fans to do a deep dive into the game and be able to look at specific player matchup comparisons and uncovered intelligence using social sentiment.

Problem Definition

We are developing a visual tool that evaluates the head to head NBA match-up of a game to predict a winner. First, we will be modeling individual player contributions to a game using several factors including historical and recent performance, social sentiment as well as matchup comparisons against opposing teams. Second, we will take the contribution of each individual player to aggregate to a total team score to predict the expected score and winner of a given game.

Literature Review

Many NBA game prediction models have been created which focus on team results and statistics (Orendorff et al (2011); Torres, R. (2013); Avalon et al (2016); Jones, E. (2016); Lin et al (2014); Zhang, Z. (2019)). Other modeling techniques focus on individual player performance using a variety of approaches including playing performance (Hongfei (2021)); Javadpour et al (2022); O'Boyle et al (2012); Beckler et al (2012); Saladi (2020)) , team chemistry (Mukherjee et al (2019)), player sentiment via interviews (Oved, N. et al (2020)) and social posts (Xu, C. et al (2014); González et al (2016), Dreyer et al (2022); Gong et al (2020)).

Team-based outcome models largely ignore player performance and contributions and player-based models largely ignore the impact of a specific matchups. There needs to be a collective approach that incorporates many of these individual elements together.

Data Sources: Data came from two sources: NBA.com and Twitter API feed.

NBA.com Statistics (via API & partially sourced from Kaggle):
Games - All games from 2004 season to end of 2023 season (April 2023)

- Data included game date, game id, home and away team ids, aggregate game stats for each team.
- Total number of fields: 21 variables (continuous and categorical)

Game Details - Box scores of each game including individual player stats:

- Data included: Game ID, player ID, Team Abbreviation, Player Name, Game Statistics (Minutes played, field goal attempted and made, points, rebounds, assists etc.)
- Total number of fields: 29 variables (continuous and categorical)

Social Sentiment via Twitter Social

Social discussion data was captured from Twitter firehouse via the Tweepy Python Library. For each player, we took a sample of fifty tweets that had the following breakdown: Tweet Text, Username, User_statuses_count, user_followers, user_location, user_verified, fav_count, rt_count(retweet count), tweet_date, and attribute_score.

Proposed Method

Our approach incorporates the following elements:

Element	Value / Benefit
Our model focuses on the starting lineup player's contribution to the total points in a given game.	Shows incremental benefit or impact of a player on a game and its predicted outcome.
We modeled individual player performance and predicted points per game based on their performance at home or away and against specific teams.	This will allow us to better understand if for a given player who is playing against a given team, will they perform above or below their expectation.
We incorporated social sentiment using Twitter data to score a player using recent social discussion.	Social sentiment offers the user a method to incorporate wisdom of the crowd and uncover potential hidden intelligence on a player.
Interactive display allows users to explore and compare different players, teams, and attributes.	Current prediction tools only show which team will play against which other team and show numbers to predict the most likely winner.

Why our approach is better

- We are placing value on the individual team member and seeking to understand how individual performance is impacted based on match-ups.
- Each final predicted score is the aggregate of many models each customized to the individual player.
- Visually, our display will make it intuitive for the user to compare potential matchups and see a breakdown of the model’s prediction.

In short, we are addressing many of the weaknesses of existing models and prediction approaches which rely heavily on team performance or a player's basic game stats.

Detailed description of your approaches: algorithms, user interfaces, etc.

Innovation #1: Player Match-up Comparison Model / Algorithm(s):

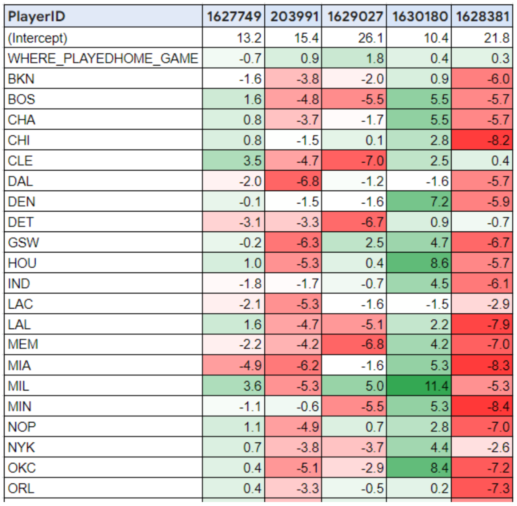
Description of how the modeling works:

For each player, we completed the following tasks:

- Ran a multiple linear regression with the players' points per game as dependent variable and all teams and all players as categorical variables using historical and recent data of the player against each team they played against
- We analyzed past 7-days of social sentiment via Twitter to create a sentiment score using a -50 to 50 score range for positive, neutral and negative
- Using combination of baseline and team/player models, we estimated points per game for each player
- For an upcoming game, we aggregated the players points for each team to determine the winner

Multiple Linear Regression Model Formula:

$$\textbf{Player}_{\textbf{A}}\textbf{Team}_{\textbf{A}}\ \textbf{Points Per Game} = \beta 0 + WhereGamePlayed\, \beta 1 + TeamPlayed\, \beta 2 + \epsilon$$



PlayerID	1627749	203991	1629027	1630180	1628381
(Intercept)	13.2	15.4	26.1	10.4	21.8
WHERE_PLAYEDHOME_GAME	-0.7	0.9	1.8	0.4	0.3
BKN	-1.6	-3.8	-2.0	0.9	-6.0
BOS	1.6	-4.8	-5.5	5.5	-5.7
CHA	0.8	-3.7	-1.7	5.5	-5.7
CHI	0.8	-1.5	0.1	2.8	-8.2
CLE	3.5	-4.7	-7.0	2.5	0.4
DAL	-2.0	-6.8	-1.2	-1.6	-5.7
DEN	-0.1	-1.5	-1.6	7.2	-5.9
DET	-3.1	-3.3	-6.7	0.9	-0.7
GSW	-0.2	-6.3	2.5	4.7	-6.7
HOU	1.0	-5.3	0.4	8.6	-5.7
IND	-1.8	-1.7	-0.7	4.5	-6.1
LAC	-2.1	-5.3	-1.6	-1.5	-2.9
LAL	1.6	-4.7	-5.1	2.2	-7.9
MEM	-2.2	-4.2	-6.8	4.2	-7.0
MIA	-4.9	-6.2	-1.6	5.3	-8.3
MIL	3.6	-5.3	5.0	11.4	-5.3
MIN	-1.1	-0.6	-5.5	5.3	-8.4
NOP	1.1	-4.9	0.7	2.8	-7.0
NYK	0.7	-3.8	-3.7	4.4	-2.6
OKC	0.4	-5.1	-2.9	8.4	-7.2
ORL	0.4	-3.3	-0.5	0.2	-7.3

Where Game Played corresponds to whether the player played at home or away (categorical)
Team Played corresponds to the team that the player is playing. It is important to note that some players will have more games played against a specific team than others. (Categorical)

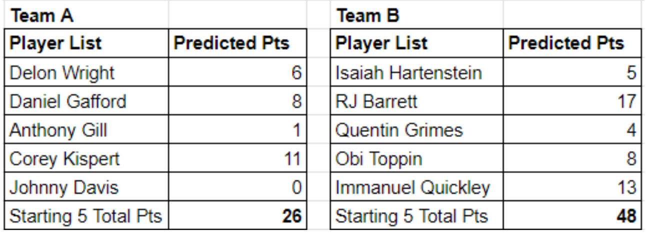
Using this modeling approach, there will be a model for each player in the NBA (~450).

The figure on the left is an example output of the coefficient for a given team with its five starting players:
Intercept shown corresponds to base level expected points per game for a given player. The matrix shows the expected points gained or loss when a player plays a particular team. This does not show p-value output but coefficients with little to no point impact tend to have high p-value.

We found that the value of being home or away is negligible in terms of points.

How to use information for prediction:

Using the model for each player, the team they are playing against and whether the game is at home or away, a prediction is determined for each of the starting five players of a given team.

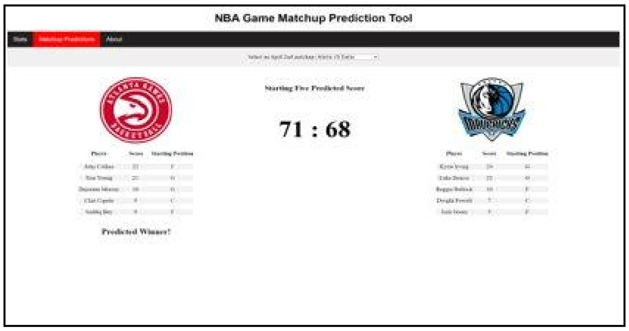


Team A			Team B	
Player List	**Predicted Pts**		**Player List**	**Predicted Pts**
Delon Wright	6		Isaiah Hartenstein	5
Daniel Gafford	8		RJ Barrett	17
Anthony Gill	1		Quentin Grimes	4
Corey Kispert	11		Obi Toppin	8
Johnny Davis	0		Immanuel Quickley	13
Starting 5 Total Pts	**26**		Starting 5 Total Pts	**48**

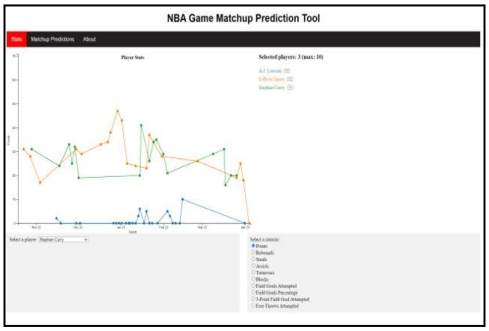
As an example, for a given game, the starting five players' models would be used to predict points given their opponent. In the figure below, Team B would be predicted to win given their starting five points prediction.

Innovation #2 Visual Tool (Developed in D3)

Our visual tool incorporates a stats section which allows users to select and compare players' game statistics such as points, rebounds, field goal attempts and many other metrics trended over time. There is also a section where users can select a game (out of 6 games) and be provided both a prediction of the game winner along with player sentiment scores as a guide.



NBA Game Matchup Prediction Tool
Stats
Matchup Predictions
About
Starting Five Predicted Score
71 : 68
Predicted Winner!



Innovation #3: Player Social Sentiment (using wisdom of the crowd):

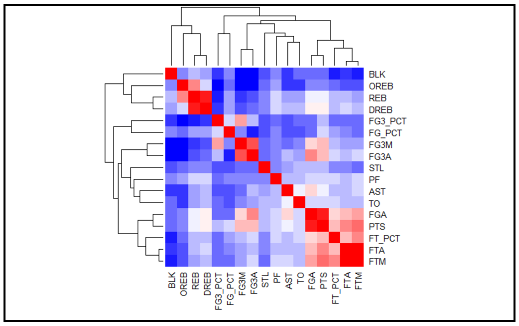
With our Twitter Firehouse we selected 50 tweets per player using a basic filtering mechanism of a baseline required level of followers, favorites, and retweets (in order to select tweets with more gravitas). After obtaining these tweets we then ran a sentiment analysis engine via Wordblob where each tweet's text would be scored on a range of -1 to 1 gradient (negative to positive sentiment with 0 being neutral). Afterwards, each score would be summed up to create a histogram on an overall aggregate spectrum of -50 to 50.

Evaluation

Description of your testbed; list of questions your experiments are designed to answer

What player metrics matter?

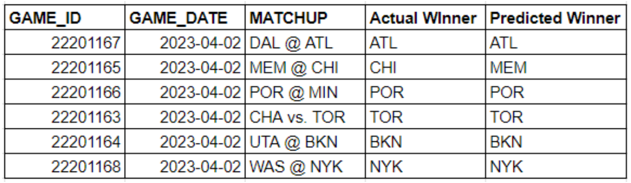
As part of the data exploratory, we performed a correlation matrix for each basketball stat in order to identify and visualize any pattern in the data. Figure below presents the correlation matrix plot. Most variables related to the act of shooting such as: FGA, FG3A, FT, among others, are



highly related to the variable of interest, points per game.Similarly, the heat map presents clearly those findings as well as the different clusters based on the correlations results. We can see that there are about 4 clusters, confirming the results of the player cluster analysis on the progress report. The 4 clusters consist of: defensive stats such as blocks and rebounds, 3-pointers, possession and shooting stats.

Since teams combine player clusters to build balanced lineups, players' performance could vary throughout season matchups. These findings helped to support using non-traditional metrics to predict output such as the team matchup.

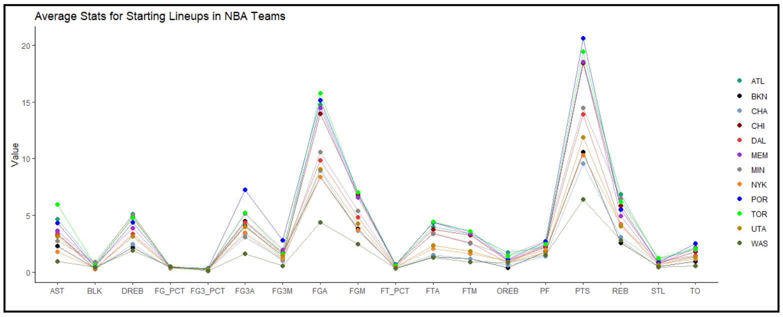
Experiment #1: Can this modeling and prediction method determine the winner?
We used our modeling process to predict the winner of games on April 2th. We focused on the starting five players of each team and aggregated the predicted points. The teams with the highest points were predicted to be the winner.



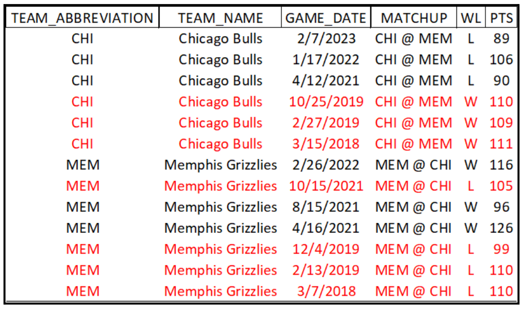
GAME_ID	GAME_DATE	MATCHUP	Actual Winner	Predicted Winner
22201167	2023-04-02	DAL @ ATL	ATL	ATL
22201165	2023-04-02	MEM @ CHI	CHI	MEM
22201166	2023-04-02	POR @ MIN	POR	POR
22201163	2023-04-02	CHA vs. TOR	TOR	TOR
22201164	2023-04-02	UTA @ BKN	BKN	BKN
22201168	2023-04-02	WAS @ NYK	NYK	NYK

In the table to the left, we can see the model predictions were correct 83% of the time. This does represent a small sample, however it is promising. The table to the left shows an example of the prediction output used to predict each team's points.

Moreover, as part of the exploratory analysis, we performed a parallel coordinates chart to visualize for the predicted teams matchups the average starting lineups (for the particular match) performance statistics. As mentioned above, we can see a clear positive relationship between FGA and PTS. Toronto Raptors starting lineup have a higher FGA, followed by Portland, which leads on Points per game.



Based on the parallel coordinates results, we can infer which team should overtake another. In addition, and more importantly, this graph helps us to determine if the performance of the players varies according to their opponent. We can see that in fact Memphis starting players have on average higher FGA and PTS per game than Chicago, which is in accordance to our predictions. Nevertheless, we failed to predict, leading Chicago to victory. The highlighted table below shows how Chicago indeed exhibits dominance against Memphis despite the overall average statistical performance for the starting lineups for the corresponding game. These



TEAM_ABBREVIATION	TEAM_NAME	GAME_DATE	MATCHUP	WL	PTS
CHI	Chicago Bulls	2/7/2023	CHI @ MEM	L	89
CHI	Chicago Bulls	1/17/2022	CHI @ MEM	L	106
CHI	Chicago Bulls	4/12/2021	CHI @ MEM	L	90
CHI	Chicago Bulls	10/25/2019	CHI @ MEM	W	110
CHI	Chicago Bulls	2/27/2019	CHI @ MEM	W	109
CHI	Chicago Bulls	3/15/2018	CHI @ MEM	W	111
MEM	Memphis Grizzlies	2/26/2022	MEM @ CHI	W	116
MEM	Memphis Grizzlies	10/15/2021	MEM @ CHI	L	105
MEM	Memphis Grizzlies	8/15/2021	MEM @ CHI	W	96
MEM	Memphis Grizzlies	4/16/2021	MEM @ CHI	W	126
MEM	Memphis Grizzlies	12/4/2019	MEM @ CHI	L	99
MEM	Memphis Grizzlies	2/13/2019	MEM @ CHI	L	110
MEM	Memphis Grizzlies	3/7/2018	MEM @ CHI	L	110

results confirm that in effect the performance of the players varies according to a particular matchup.

Experiment #2: Player Points Prediction

We tested 10 players to predict their points per game using the model for the games occurring after January 1, 2023. Our per game points prediction for each of the 10 players tested yields an Mean Squared Error $(\text{Prediction - Actual})^2$ in the range of 33-300 (average - 97.62).

Experiment #3: Social Sentiment

Data collection proved to be challenging given recent changes at Twitter, however, we found that sampling 50 tweets and assigning scores of -1 to negative posts, 0 to neutral and 1 to positive messages to create an aggregate score proved to be effective at giving a sense of potential issues or challenges that may exist with a player, however, its in a model was limited due to some of the data collection issues. At the outset the Twitter data was limited to at most a week back in history and the overall tweet output was rate limited. Toward the end of our implementation Twitter API access was completely revamped due to new company policy changes that further restricted access for free usage. As a result, the group determined that social sentiment would be used as a complimentary metric instead. With unlimited resources we would probably like to purchase more of this data to access further back in history as well as increase our sample size from 50 tweets to perhaps more than a thousand instead which would remove possible volatility/sample bias.

Conclusions and Discussion

Using a player driven approach where a given player's historical performance against other teams as well as their performance home or away is modeled and used to predict their point scored against a given team proved to be effective. When aggregating the starting five lineup of one team versus another, our approach proved to predict 83% of games in a limited sample. In addition, when predicting player points per game and comparing it to actual results saw reasonable MSE results.

Social sentiment proved to be a challenge in terms of data collection driven by instability at Twitter. Once collected, it was used as a complimentary metric that the users of the tool may use to determine if there is a positive or negative value of a particular team member.

Potential weaknesses of the model include teams with unstable lineups or when younger players are used with limited historical performance against teams. This is likely to occur later in the season when certain teams might be incentivized to lose.

Overall, this tool can be helpful for bettors as well as coaches to better understand expected player performance against specific teams.

Team Members Contribution to Project

All team members contributed equally to the project.

References

Avalon, G., Balci, B., Guzman, J. (2016), Various Machine Learning Approaches to Predicting NBA Score Margins, http://cs229.stanford.edu/proj2016/report/Avalon_balci_guzman_various_ml_approaches_NBA_Scores_report.pdf

Beckler, M., Papamichael, M., Wang, H. (2009), NBA Oracle, https://www.mbeckler.org/coursework/2008-2009/10701_report.pdf

D., Johnson, T., (2011). First-Order Probabilistic Models for Predicting the Winners of Professional Basketball Games, https://www.researchgate.net/publication/228742804_First-Order_Probabilistic_Models_for_Predicting_the_Winners_of_Professional_Basketball_Games

Dreyer, F., Greif, J., Gunther, K., Spiliopoulou, M., Niemann, U. (2022) Data-Driven Prediction of Athletes' Performance based on their Social Media Presence, https://uliniemann.com/pdf/ds-2022-preprint.pdf

ESPN, (2022). Americans could bet $3.1 billion on NCAA men's basketball tournament, according to survey, https://www.espn.com/chalk/story/_/id/33500830/americans-bet-31-billion-ncaa-men-basketball-tournament-according-survey

Feddersen, A., Humphreys, B., Soebbing, B. (September 2013) Sentiment Bias and Asset Prices: Evidence from Sports Betting Markets and Social Media, http://busecon.wvu.edu/phd_economics/pdf/13-07.pdf

González, C., García-Nieto, J. Navas-Delgado, I., Aldana-Montes, J. (March 2016). A Fine Grain Sentiment Analysis with Semantics in Tweets. https://idus.us.es/bitstream/handle/11441/108366/1/A%20Fine%20Grain%20Sentiment%20Analysis.pdf?sequence=1

Gong, H., Watanabe, N., Soebbing, B., Brown, M., Nagel, M. (2021). Do Consumer Perceptions of Tanking Impact Attendance at National Basketball Association Games? A Sentiment Analysis Approach, https://scholarship.rice.edu/bitstream/handle/1911/110703/JSM-2020-0274_online.pdf?sequence=1

Hongfei, L.,Maolin, Z. (2021) Artificial Intelligence and Neural Network-Based Shooting Accuracy Prediction Analysis in Basketball, Mobile Information Systems, vol. 2021, Article ID 4485589, 11 pages, 2021. https://doi.org/10.1155/2021/4485589

Hongfei, L., Maolin, Z. (June 2021) AI and Edge Computing-Driven Technologies for Knowledge Defined Networking, Mobile Information Systems, Article ID 4485589, https://doi.org/10.1155/2021/4485589

Javadpour, L., Blakeslee, J., Khazaeli, M., Schroeder, P. (March 2022) Optimizing the best play in basketball using deep learning, Journal of Sports Analytics, vol. 8, no. 1, pp. 1-7, 2022, https://content.iospress.com/articles/journal-of-sports-analytics/jsa200524#:~:text=More%20recently%2C%20data%20analytics%20in,et%20al.%2C%202006).

Jones, E. (April 2016) Predicting Outcomes of NBA Basketball Games, https://library.ndsu.edu/ir/bitstream/handle/10365/28084/Predicting%20Outcomes%20of%20NBA%20Basketball%20Games.pdf

Koster J, Aven B (2018) The effects of individual status and group performance on network ties among teammates in the National Basketball Association. PLoS ONE 13(4): e0196013. https://doi.org/10.1371/journal.pone.0196013

Lin, J., Short, L., Sundaresan, V. (2014) Predicting National Basketball Association Winners http://cs229.stanford.edu/proj2014/Jasper%20Lin,%20Logan%20Short,%20Vishnu%20Sundaresan,%20Predicting%20National%20Basketball%20Association%20Game%20Winners.pdf

Mukherjee, S., Huang, Y., Neidhardt, J. *et al.* (2019) Prior shared success predicts victory in team competitions. *Nat Hum Behav* 3, 74–81 . https://doi.org/10.1038/s41562-018-0460-y

Murakami-Moses, M. (2020) Analysis of Machine Learning Models Predicting Basketball Shot Success, https://www.the-iyrc.org/uploads/1/2/9/7/129787256/20_iyrc2020_35_final.pdf

O'Boyle, E., Aguinis, H. (February, 2012). The Best And The Rest: Revisiting The Norm of Normality Of Individual Performance, https://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.2011.01239.x

Oved, N., Feder, A., Reichart, R. (September 2020). Predicting In-Game Actions from Interviews of NBA Players, Computational Linguistics (2020) 46 (3): 667–712 https://direct.mit.edu/coli/article/46/3/667/93377/Predicting-In-Game-Actions-from-Interviews-of-NBA

Torres, R. (December, 2013) Prediction of NBA games based on Machine Learning Methods, https://paperzz.com/doc/7255157/prediction-of-nba-games-based-on-machine-learning-methods

Xu, C., Yang, Y., Hoi, C. (2014). Rolling Moneyball with Sentiment Analysis: What do NBA Players' Tweets Tell?
https://www.researchgate.net/profile/Yang-Yu-219/publication/272826079_Rolling_Moneyball_wi

th_Sentiment_Analysis-What_do_NBA_Players'_Tweets_Tell/links/56538bd108aefe619b196210/Rolling-Moneyball-with-Sentiment-Analysis-What-do-NBA-Players-Tweets-Tell.pdf

Zhang, X. (December, 2019) Modeling of NBA Game Data and their Correlation Structure, https://core.ac.uk/download/pdf/270201029.pdf

Zhuo, S., Mingrui, L., Meng, W., Jing, S., Wei, C., Xiaonan, L., (2022) NPIPVis: A Visualization System Involving NBA Visual Analysis and Integrated Learning Model Prediction, Virtual Reality & Intelligent Hardware, Volume 4, Issue 5, Pages 444-458, https://doi.org/10.1016/j.vrih.2022.08.008