競輪予想記事の自動生成に向けた深層学習によるレース結果予測

Race Results Prediction Using Deep Learning for Generation of Predicted Article in Keirin

吉田拓海[1*]　横山想一郎[1]　山下倫央[1]　川村秀憲[1]
Takumi Yoshida[1], Soichiro Yokoyama[1], Tomohisa Yamashita[1], and Hidenori Kawamura[1]

[1] 北海道大学大学院情報科学研究科
[1] Graduate School of Information Science and Technology, Hokkaido University

Abstract: Proactive information dissemination is necessary to revitalize the Japanese Keirin business. An article describing prediction of the race result is effective for revitalizing the Keirin business. However, writing articles manually is a heavy burden. This is because the number of races per day is about 60 races and it is very large. Another reason is that participating racers will be decided according to the race result of the previous day. Therefore, there is a strong demand for a technique for automatically generating content including articles. The purpose of this research is to generate articles automatically in Japanese Keirin. In order to automatically generate articles, we predict race results by machine learning. And, we generate an article explaining the predicted results using the template.

1　はじめに

スポーツの分野においては，記事の自動生成に関する研究が積極的に行われている．野球においては，打者成績からイニング速報を生成する研究 [1] や，テキスト速報からイニングの要約文を生成する研究 [2] が行われている．また，スポーツ以外の分野においても，時系列数値データから概況テキストを生成する研究 [3] や，天気予報コメントを自動生成する研究 [4] が行われる．上記の研究は，手動でテキストを生成する負担を軽減するために行われている．

競輪の 1 日あたりのレース数は約 60 レースあり，その各レースに予想記事が手動で生成されている．本稿では”予想記事”という用語を，レース結果の予測を記述する記事という意味で用いる．また，前日のレース結果に応じて参加選手が決定されるため，記事の生成の負担が大きい．一方で，競輪業界では新規ユーザの獲得が課題となっており，積極的な情報発信が求められている．

したがって，本研究の目的は，競輪における予想記事の自動生成である．深層学習によってレース結果を予測し，予測したレース結果から上位予想選手を選択する．本稿では，レース結果として 1 着から 3 着の選手を予測し，各選手を上位予想選手として選択する．記事は予め用意したテンプレートに対し上位予想選手や，上位予想選手の過去成績，上位予想選手のライン構成などの条件に応じたテキストを当てはめることによって生成する．

本稿では，第 2 章，第 3 章で本研究で取り扱う競輪について示す．第 4 章でレース結果予測のモデル，比較実験について示し，第 5 章で記事生成について詳細を示す．最後に第 6 章でまとめと今後の展望について示す．

*連絡先：北海道大学大学院情報科学研究科
〒 060-0814 札幌市北区北 14 条西 9 丁目
E-mail: yoshida@complex.ist.hokudai.ac.jp

2　競輪

競輪は，自転車を使用した日本発祥のトラックレースである．レースは基本的に 9 人で行い，バンクと呼ばれる競争路を周回し，ゴールを競う．また，選手はレース中にラインと呼ばれる縦列を形成する．多くの場合は，同じ地域に所属する選手同士でラインを組む．ラインの先頭を走る選手 (先行選手) は，走る速度や加速のタイミングを自由に決定することができるが，風の抵抗をラインの中で最も受けるため，他の選手よりも体力を消耗する．先行選手の後ろを走る選手 (番手選手) は，前方の選手を風よけに使用することができるため，体力を温存することができる．その代わり，番手選手は後ろから迫ってくる他ラインの選手をブロックすることで，先行選手を援護する役割がある．選手は，最後のゴール前の直線まではチーム (ライン) で走るが，最後は各個人が 1 位を目指す．このような，ラインに

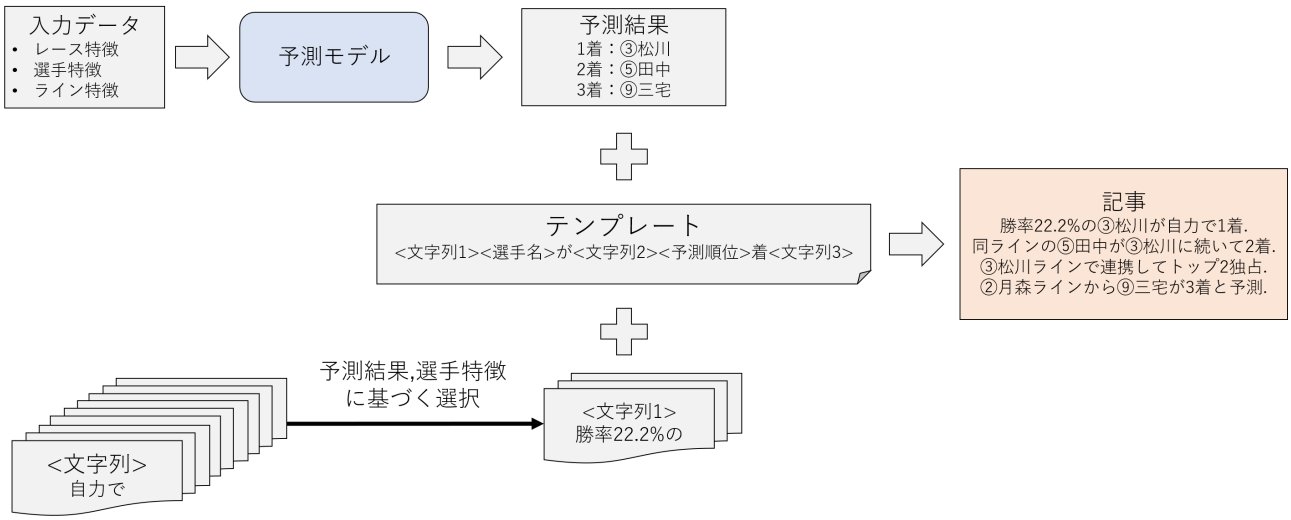

図 1: 記事生成の処理の概要

よって作られるレース展開は，競輪特有のものである．

競輪とよく似た競技として競馬が挙げられる．競馬に関しては，レース結果を予測する研究が幾つか行われている．ニューラルネットワークを用いた研究 [5] や，サポートベクターマシンを用いた研究 [6]，ロジスティック回帰を用いた研究 [7]，ファジィ論理を用いた研究 [8] などが行われている．競輪と競馬との違いは，馬か自転車か，出場人数，開催頻度などあるが，最大の違いはラインである．ラインによる選手間の相互作用を考慮することは，競輪の予測において重要である．

3 競輪における予想記事

本章では，現在人手によって作成されている予想記事にどのような情報が含まれているかを検証する．さらに，どのように予想記事を自動生成するかを説明する．

3.1 既存記事

現在，競輪では実施される全レースに対して人手によって予想記事が作成され，新聞や投票サイトでユーザに向けて配信されている．投票サイトで配信されている既存記事の例を以下に示す．既存記事は，オッズパーク [9] より引用した．

- 磯島が駆けて番手の野木が本命．鋭さ光るのは丸山だ．逆転の捲りに一考．目標の高鍋次第で大久保、攻め多彩な吉田も怖い．
- 持ち前のスピードを全開なら⑤松本―②蓮井でワンツー．自在③田中―⑨米原の捲りや同県⑥山田にマークする④竹内や⑦内村も単級
- 藤縄ー中澤の近畿勢に期待．隅田ー高津，中部勢の抵抗も見ものだろう．
- 函館最終戦のオープニングレースは２分の争いとなった．ラインの先頭を走る自力型の先行力は，ほぼ互角とみるが１０９期のルーキー神山が主導権を握ると見て番手を回る①田中が鋭く追い込む．ラインの⑤松本が続き本線を形成．対し大阪勢も侮れない．ライン４車の利を活かし④斎藤が巧く駆けると番手を回る②陶器とのワンツーも狙える．

既存記事が含む情報として次の 4 つが挙げられる．

- 上位予想選手
- 上位予想選手に関する情報
- ラインに関する情報
- レース展開に関する情報

また，既存記事には次の 2 つの特徴が挙げられる．

- ライン決着の予想
- 複数のレース展開予想

ラインによって生じる選手間の相互作用が，レース結果に大きな影響を与えているため，ラインによる決着の予想が既存記事の特徴となっていると考えられる．ラインの影響により，レースの流れによってレース結果が大きく異なることが想定されるため，既存記事には複数のレース展開の予測が記載されている．以上のことからも，競輪では，ラインが重要であることがわかる．

3.2 記事生成アプローチ

記事生成アプローチの流れ図を図 1 に示す

本稿では先述した 4 つの情報の内，現時点でデータのないレース展開に関する情報を除いた，次の 3 つの情報を含む記事を生成することを目的とする．

- 上位予想選手
- 上位予想選手に関する情報
- ラインに関する情報

機械学習によってレース結果を予測し，予測したレース結果から上位予想選手を選択する．本稿では，レース結果として 1 着から 3 着の選手を予測し，各選手を上位予想選手として選択する．記事は予め用意したテンプレートに対し上位予想選手や，上位予想選手の過去成績，上位予想選手のライン構成などの条件に応じたテキストを当てはめることによって生成する．

4 レース結果予測

本章では，予想記事生成のために，レース結果を機械学習を用いて予測する手法を提案して，その性能の比較検証を行う．予測するレース結果は，1 着から 3 着となる選手の順序である．

4.1 予測アプローチ

予測モデルの入出力の設定について説明する．上位 3 人の順序を予測するための問題設定として，次の 3 つが考えられる．1 つ目は，各選手の強さを 1 人ずつ推定するポイントワイズアプローチである．しかし先述した通り，競輪はラインによって生じる選手間の相互作用がレース結果に大きな影響を及ぼすため，選手 1 人のみに注目して予測を行う手法は，適切ではないと考えられる．2 つ目は，選手 2 人の順序関係を学習するペアワイズアプローチ (2 人モデル) である．3 つ目は，選手 9 人の順序関係を直接学習するリストワイズアプローチ (9 人モデル) である．本稿ではこの 2 人モデルと 9 人モデルによって予測を行い，その性能を比較する．

4.1.1 2 人モデル

2 人モデルは選手 2 人の順序関係を学習するモデルである．レースの情報と選手 2 人の情報を二値分類器に入力した出力値を 1 レース分集計し，それらの値から上位選手の順序確率を近似的に計算する．以降では，二値分類器の入出力と順序確率の計算について説明する．

入力/出力　2 人モデルは選手 2 人 (車番 i，車番 j) の情報を入力とし，$x_{i,j}$ と表記する．入力 $x_{i,j}$ の教師ラベル $t_{i,j}$ を次のように定義する．

$$t_{i,j} = \begin{cases} 0 & place_i > place_j \\ 1 & place_i < place_j \end{cases} \tag{1}$$

ここで，$place_i$ は車番 i の順位である．$x_{i,j}$ の出力 $y_{i,j}$ は，車番 i が車番 j より上位になる確率として扱う．

順序確率の計算　二値分類器の出力値から上位選手 (1,2,3 位) の順序確率を計算する．ここで，出力値は $y_{i,j} \neq 1 - y_{j,i}$ となっているため，集計時に式 2 のような補正を行う．

$$y_{i,j}(\text{補正後}) = \frac{y_{i,j} + (1 - y_{j,i})}{2} \tag{2}$$

式 2 の補正によって $y_{i,j} = 1 - y_{j,i}$ が成立する．補正後の出力値を用いて順序確率を近似的に計算する．次の 3 つの確率をそれぞれ計算し，その積によって順序確率を計算する．

- 車番 a が 1 着になる確率
- 車番 a が 1 着の時，車番 b が 2 着になる確率
- 車番 a が 1 着，車番 b が 2 着の時，車番 c が 3 着になる確率

各確率の計算式を次に示す．

$$p(1\text{ 着} = a) = \frac{\prod_{j \neq a} y_{a,j}}{\sum_i (\prod_{j \neq i} y_{i,j})} \tag{3}$$

$$p(2\text{ 着} = b | 1\text{ 着} = a) = \frac{\prod_{j \neq a,b} y_{a,j}}{\sum_i (\prod_{j \neq a,i} y_{i,j})} \tag{4}$$

$$p(3\text{ 着} = c | 1\text{ 着} = a, 2\text{ 着} = b) = \frac{\prod_{j \neq a,b,c} y_{a,j}}{\sum_i (\prod_{j \neq a,b,i} y_{i,j})} \tag{5}$$

式 3,4,5 の積を順序 $a - b - c$ の順序確率として扱う．

4.1.2 9 人モデル

9 人モデルは，選手 9 人を比較することによってレース結果を予測するモデルである．このモデルは，2 人モデルのような処理は行わずに上位 3 人の順序を直接学習する．

入力/出力　9 人モデルは選手 9 人 (車番 1, 車番 2,..., 車番 9) の情報を入力とする．対応する選手のない入力値は 0 とする．このモデルは，上位 3 人の順序確率を出力する．本稿では，9 人モデルについて，再帰型ニューラルネットワーク (RNN) による予測を行う．

RNN　RNN は系列データの学習に用いられるニューラルネットワークであり，音声，自然言語，動画像の分野で高い性能を示している．また，階層的なラベルの分類に RNN を使用する研究 [10] も行われており，RNN を用いることで高い精度を示している．本研究においても，1 着選手 (9 通り),1 着選手と 2 着選手 (72 通り), 1 着選手と 2 着選手と 3 着選手 (504 通り) の 3 つが階層的な構造になってることに注目し，RNN による予測を行う．RNN のネットワーク構造の概略図を図 2 に示す．

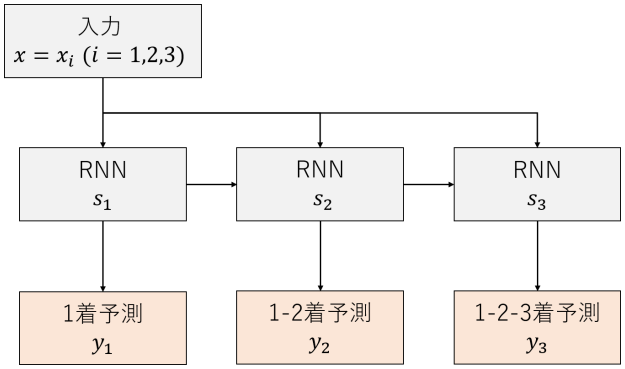

図 2: RNN の概略図

4.2　実験

前節で提案した予測手法の性能を比較するために，各手法の出力する上位 2 人と上位 3 人の順序的中率を比較する．

4.2.1　実験設定

データ　2014 年から 2016 年の間に実施されたレースのデータを使用した．2014 年と 2015 年のデータを訓練データとして使用し，2016 年のデータはテストデータとして使用した．各レース数は訓練データが 40,230 レースで，テストデータが 15,731 レースである．本稿では，以下の条件を満たすデータを使用した．

- 男性レース
- 競り無し
- 同着無し
- 欠損値無し

特徴量　機械学習では，使用する入力特徴量は予測性能に大きな影響を与える．本稿では，インターネット上で誰でも閲覧可能なデータを入力特徴量に使用する．特徴量の詳細を表 1 に示す．

車番と年齢以外の選手特徴は過去 4ヶ月のデータである．ライン特徴は，ラインの影響を予測するための特徴であり，各モデルはそのモデルに適した異なるライン特徴を使用する．2 人モデルは，先行選手の選手特徴と，比較する 2 人の選手が同じラインか否かを表す特徴を使用する．9 人モデルは，自身のラインを表すラインベクトルをライン特徴を使用する．ラインが 1-9-3 の場合，ラインベクトルは，$[1,0,1,0,0,0,0,0,1]$ となる．2 人モデルの入力は 64 次元，9 人モデルの入力は 219 次元である．

表 1: 入力特徴量

レース特徴	レースの格
	トラックの長さ
	出走人数
選手特徴	車番
	年齢
	競走得点
	1 着回数
	2 着回数
	3 着回数
	1 着率
	2 連対率
	バック取り回数
	決まり手回数
ライン特徴	先行選手か否か
	ライン内での自身の位置
	先行選手 (2 人モデル)
	ラインが同じか否か (2 人モデル)
	ラインベクトル

比較　2 人モデルと 9 人モデルを比較する．また，各モデルの学習アルゴリズムとして，ランダムフォレスト，MLP，RNN の 3 つを使用した．訓練データの 2 割を検証データとして，各モデルのハイパーパラメータを決定した．比較項目として，上位 3 人の順序的中率と上位 2 人の順序的中率，それぞれの top-k-accuracy を使用した．

4.2.2　精度比較 結果/考察

各モデルの上位 3 人の順序と上位 2 人の順序についての top-k-accuracy(k=1,5,10) を表 2 に示す．2 人モデルよりも 9 人モデルの方が精度が高い．2 人モデル

表 2: 予測モデルの比較

	上位 2 人@1 [%]	上位 2 人@5 [%]	上位 2 人@10 [%]	上位 3 人@1 [%]	上位 3 人@5 [%]	上位 3 人@10 [%]
2 人モデル (RF)	14.1	41.0	57.1	4.1	15.0	23.8
2 人モデル (MLP)	15.4	43.4	59.4	4.9	17.1	26.5
9 人モデル (RF)	15.1	40.6	55.6	4.4	14.5	22.5
9 人モデル (MLP)	17.1	45.7	61.9	5.8	18.6	27.6
9 人モデル (RNN)	18.1	47.2	63.2	6.0	19.7	29.1
オッズ	21.2	52.6	68.9	9.1	26.2	37.2

の入力は比較する 2 人の選手と，各選手の先行選手の情報しか含まれていない．それによってラインに関する情報が不足していたことが，2 人モデルが 9 人モデルより精度が低い原因であると考えられる．また，学習アルゴリズムについて注目すると，RNN によるモデルが最も高い精度を示した．

しかし，オッズの予測精度は，上位 2 人@1 が 21.2%，上位 3 人@1 が 9.1%であり，現時点ではオッズの予測に精度の面では敵わない．精度向上のための今後の課題として，有効な入力特徴量の探索が挙げられる．

5 予想記事の生成

本章では，前章で得られたレースの予測結果を用いて予想記事を生成するプロセスを説明する．

5.1 記事生成の流れ

図 1 に示した様な流れで予想記事を生成する．予測したレース結果と選手の情報から，テンプレートに当てはめる文を選択し，選択された文をテンプレートに当てはめることによって記事を生成する．以降では，テンプレートとテンプレートに当てはめる文の選択について説明する．

5.1.1 テンプレート

設定したテンプレートを次に示す．

- ＜文字列 1＞＜選手名＞が＜文字列 2＞＜予測順位＞着＜文字列 3＞

このテンプレートに任意の文字列を当てはめることによって予想記事を生成する．テンプレートに実際に文字列を当てはめると次のようになる．

- 勝率 22.2%の③松川が自力で 1 着．
  - ＜文字列 1＞：勝率 22.2%の
  - ＜選手名＞：③松川
  - ＜文字列 2＞：自力で
  - ＜予測順位＞：1
  - ＜文字列 3＞：．

このように，文テンプレートに対して，説明したい状況に応じた＜文字列＞を当てはめることによって，その状況を説明する文が生成可能である．

5.1.2 予測結果に基づく文字列の選択

予測したレース結果のライン構成に応じて文を選択することで，ラインの情報を含む記事を生成する．予測した 1 位と 2 位と 3 位が同じラインの選手であった場合には，それを説明する記事を生成することが目的である．ライン構成は，上位 3 人の選手の所属するラインと上位 3 人の選手のラインでの位置 (単騎，先行，番手) によって決定され，合計で 58 通り存在する．

5.1.3 選手情報に基づく文字列の選択

選手情報に応じて文字列を選択することによって，選手情報を含む記事を生成する．以下のデータを選手情報として出走表から抽出する．

- 過去 4ヶ月の 1 着回数
- 過去 4ヶ月の勝率
- 過去 4ヶ月の 2 連対率
- 年齢

抽出したデータに関して条件を設定し，その条件に応じて修飾文を選択する．以下の条件を設定した．

- 過去 4ヶ月の 1 着回数 $\geqq 10$
- 過去 4ヶ月の勝率 $\geqq 0.2$
- 過去 4ヶ月の 2 連対率 $\leqq 0.4$
- 年齢 $\leqq 25$
- 年齢 $\geqq 40$

5.2 生成記事結果・考察

実際に生成された記事の例と既存記事を表 3,4,5 に示す．既存記事はオッズパーク [9] より引用した．第 3.2 項で示した次の 3 つの情報を含む記事が生成されていることがわかる．

- 上位予想選手
- 上位予想選手に関する情報
- ラインに関する情報

ライン構成を説明する記事が生成できていることが確認できる．例 1 は，1 位の選手と 2 位の選手が同じラインに所属していることを説明している．同様に例 2 は，1 位から 3 位の選手が全員同じラインに所属していることを説明している．また，上位予想選手に関する情報を含む記事を生成できていることが確認できる．例 1 では，1 位になると予測した選手の過去 4ヶ月の勝率が 22.2%であるという記述がされている．

既存記事と比較すると，生成記事には既存記事と同じ選手が記載されていることが多い．例 5 では，生成記事に記載されている滝本，戸伏，松尾の 3 人が既存記事にも有力選手として記載されている．これは，生成された記事の予想内容が悪くないことを示している．生成記事の課題としては，既存記事と比較して語彙や文のパターンが乏しいこと，複数のレース展開予測が記述できていないことが挙げれられる．

表 3: 生成例 1

	2018-5-23 向日町競輪場 第 9 レース
既存記事	意地見せる地元高久保，器用な月森相手だが，松川が一枚上手か.
生成記事	勝率 22.2%の③松川が自力で 1 着. 同ラインの⑤田中が③松川に続いて 2 着. ③松川ラインで連携してトップ 2 独占. ②月森ラインから⑨三宅が 3 着と予測.

表 4: 生成例 2

	2018-5-23 岐阜競輪場 第 4 レース
既存記事	二日目余裕で逃げ切った⑤山本のパワーが断トツ。ここも力任せの先行で豪快に逃げ切る。近畿勢⑦辻本が喰い下がると見るも、③安藤や①山中も気力次第か。
生成記事	勝率 33.3%の⑤山本が自分の脚で 1 着. 同ラインの⑦辻本が⑤山本に続いて 2 着. ②谷が 3 着. ⑤山本ラインが別線を抑えて上位独占と予測.

表 5: 生成例 3

	2018-5-23 佐世保競輪場 第 7 レース
既存記事	互角戦だが⑦滝本の自力に期待．①戸伏の逆転も．④内藤の連入十分．②飯塚も自力で単も．穴は⑥松尾の抜け出しから.
生成記事	二連対率 45.8%の⑦滝本が自力で 1 着. 同ラインの番手①戸伏が⑦滝本に続いて 2 着. ⑦滝本ラインでトップ 2 独占. 別線の⑥松尾が 3 着と予測.

6 まとめと今後の展望

本稿では，機械学習によって競輪のレース結果を予測し，予測結果に基づく記事を生成した．選手 2 人を比較するモデル (2 人モデル) と，選手 9 人を比較するモデル (9 人モデル) を比較し，9 人モデルの有効性を示した．また，RNN を使用することによる精度向上を確認した．予測結果に基づき，テンプレートを用いて予測結果を説明する記事を生成した．

今後の課題として，レース結果予測についてはオッズよりも高い精度を示す予測モデルを設計することが挙げられる．記事生成については，語彙や文パターンが豊富な記事の生成と，複数のレース展開の予測を記述する記事の生成が今後の課題である．

謝辞

本研究は，株式会社チャリ・ロトの支援を受け実施されたものです．ここに感謝の意を表します．

参考文献

[1] 村上聡一郎, 笹野遼平, 高村大也, 奥村学. 打者成績からのイニング速報の自動生成. 言語処理学会第 22 回年次大会発表論文集, 2016.

[2] 田川裕輝, 嶋田和孝. テンプレートの自動生成によるイニングの要約文生成. 言語処理学会第 23 回年次大会発表論文集, 2017.

[3] 村上聡一朗, 渡邉亮彦, 宮澤彬, 五島圭一, 柳瀬利彦, 高村大也, 宮尾祐介. 時系列数値データからの概況テキストの自動生成. 言語処理学会第 23 回年次大会発表論文集, 2017.

[4] 村上聡一朗, 笹野遼平, 高村大也, 奥村学. 数値予報マップからの天気予報コメントの自動生成. 言語処理学会第 23 回年次大会発表論文集, 2017.

[5] Elnaz Davoodi and Ali Reza Khanteymoori. Horse racing prediction using artificial neural networks. In *Proceedings of the 11th WSEAS International Conference on Nural Networks and 11th WSEAS International Conference on Evolutionary Computing and 11th WSEAS International Conference on Fuzzy Systems*, NN'10/EC'10/FS'10, pp. 155–160, Stevens

Point, Wisconsin, USA, 2010. World Scientific and Engineering Academy and Society (WSEAS).

[6] David Edelman. Adapting support vector machine methods for horserace odds prediction. *Annals of Operations Research*, Vol. 151, No. 1, p. 325, Nov 2006.

[7] Noah Silverman and Marc Suchard. Predicting horse race winners through a regularized conditional logistic regression with frailty. *Journal of Prediction Markets*, Vol. 7, No. 1, pp. 43–52, 2013.

[8] Manish Jogeeah, Akshay Kumar Chandoo, Selukoto Paupiah, and Sameerchand Pudaruth. Using fuzzy logic to predict winners in horseraces at the champ de mars. 01 2015.

[9] オッズ・パーク株式会社. 競輪（keirin・ケイリン）ならオッズパーク競輪｜予想情報も充実！ `http://www.oddspark.com/keirin/`. (Accessed on 05/29/2018).

[10] Yanming Guo, Yu Liu, Erwin M. Bakker, Yuanhao Guo, and Michael S. Lew. Cnn-rnn: a large-scale hierarchical image classification framework. *Multimedia Tools and Applications*, Vol. 77, No. 8, pp. 10251–10271, Apr 2018.