

The assignment success for 22 horse breeds registered in the Czech Republic: The machine learning perspective

LENKA ŠTOHLOVÁ PUTNOVÁ^{1*}, RADEK ŠTOHL²

¹Department of Animal Morphology, Physiology and Genetics, Faculty of AgriScience, Mendel University in Brno, Brno, Czech Republic

²Department of Control and Instrumentation, Faculty of Electrical Engineering and Communication, University of Technology, Brno, Czech Republic

*Corresponding author: putnova@email.cz

The authors are fully responsible for both the content and the formal aspects of the electronic supplementary material. No editorial adjustments were made.

Electronic Supplementary Material (ESM)

Table S1. The Reynolds' D_R genetic distances (below the diagonal; $n = 7\,588$) and paired F_{ST} values (above the diagonal; $n = 1\,742$) among the Czech-registered breeds

Table S2. The individual assignment success rates for each horse breed, achieved by using eight different assignment methods in GENECLASS

Table S3. The confusion matrix of the Bayesian method (the Rannala and Mountain approach), as calculated by GENECLASS for each Czech-registered horse breed ($n = 1\,742$; accuracy 83.6%)

Table S4. The confusion matrix of the Bayes Network classifier, as calculated by WEKA for each Czech-registered horse breed ($n = 1\,742$; accuracy 78.4%)

Table S5. The confusion matrix of the Bayesian clustering method, as calculated by STRUCTURE for each Czech-registered horse breed ($n = 1\,742$; accuracy 58.03%)

Abbreviations of horse breeds used in tables: ARAB = Arabian; CMB = Czech-Moravian Belgian Horse; CSP = Czech Sport Pony; CZW = Czech Warmblood; HAF = Haflinger; HUC = Hucul; IRI = Irish Cob; KIN = Kinsky Horse; KLA = Old Kladruber Horse; LIP = Lipizzan; MIN = Miniature Horse; MOW = Moravian Warmblood; NOR = Noriker; SHA = Shagya; SHP = Shetland Pony; SLW = Slovak Warmblood in the Czech environment; SNOR = Silesian Noriker; STA = Standardbred; THO = Thoroughbred; TRA = Trakehner; WPB = Welsh Part Bred; WPC = Welsh Pony and Cob

Table S1. The Reynolds' D_R genetic distances (below the diagonal; $n = 7\ 588$) and paired F_{ST} values (above the diagonal; $n = 1\ 742$) among the Czech-registered breeds

	ARAB	CMB	CSP	CZW	HAF	HUC	IRI	KIN	KLA	LIP	MIN	MOW	NOR	SHA	SHP	SLW	SNOR	STA	THO	TRA	WPB	WPC
ARAB	0.1299	0.0542	0.0559	0.1288	0.1038	0.1082	0.1082	0.0627	0.0797	0.0862	0.1139	0.0704	0.1076	0.0500	0.1441	0.0588	0.1167	0.0938	0.0841	0.0551	0.0508	0.0989
CMB	0.1330	0.0716	0.0780	0.0615	0.0946	0.0711	0.0904	0.0939	0.1385	0.0947	0.1233	0.0567	0.1388	0.1294	0.0774	0.0646	0.1275	0.1241	0.0848	0.0640	0.0708	
CSP	0.0597	0.0758	0.0297	0.0732	0.0505	0.0473	0.0398	0.0597	0.0698	0.0505	0.0592	0.0550	0.0590	0.0705	0.0281	0.0671	0.0761	0.0606	0.0397	0.0080	0.0328	
CZW	0.0584	0.0805	0.0335	0.1010	0.0727	0.0604	0.0087	0.0668	0.0812	0.0761	0.0288	0.0693	0.0735	0.0973	-0.0010	0.0836	0.0562	0.0160	0.0059	0.0254	0.0490	
HAF	0.1360	0.0708	0.0805	0.1058	0.0918	0.0910	0.1099	0.0847	0.1337	0.1164	0.1219	0.0586	0.1258	0.1434	0.0979	0.0627	0.1336	0.1438	0.1128	0.0681	0.0801	
HUC	0.1113	0.1002	0.0569	0.0771	0.0941	0.0938	0.0778	0.0998	0.0956	0.0914	0.0859	0.0807	0.0954	0.1068	0.0666	0.0908	0.0892	0.1004	0.0832	0.0516	0.0645	
IRI	0.1149	0.0768	0.0564	0.0683	0.1009	0.1021	0.0809	0.0982	0.1160	0.0698	0.1021	0.0570	0.1043	0.0970	0.0610	0.0720	0.1139	0.1074	0.0769	0.0454	0.0433	
KIN	0.0677	0.0956	0.0454	0.0112	0.1180	0.0878	0.0886	0.0737	0.0844	0.0932	0.0283	0.0853	0.0780	0.1117	0.0084	0.0975	0.0573	0.0211	0.0121	0.0326	0.0585	
KLA	0.0874	0.0968	0.0665	0.0747	0.0860	0.1063	0.1103	0.0821	0.1073	0.0933	0.1075	0.0804	0.1003	0.1276	0.0677	0.0911	0.1097	0.1052	0.0754	0.0386	0.0659	
LIP	0.1155	0.1623	0.1017	0.1096	0.1640	0.1288	0.1485	0.1147	0.1410	0.1289	0.0854	0.1017	0.1106	0.1395	0.0763	0.1117	0.1004	0.1028	0.0731	0.0765	0.1018	
MIN	0.1245	0.1041	0.0644	0.0870	0.1250	0.1008	0.0854	0.1053	0.1109	0.1632	0.1110	0.0671	0.1199	0.0157	0.0739	0.0862	0.1380	0.1225	0.0881	0.0520	0.0545	
MOW	0.0791	0.1308	0.0697	0.0358	0.1307	0.0999	0.1141	0.0379	0.1147	0.1192	0.1270	0.0938	0.0841	0.1237	0.0269	0.1107	0.0544	0.0332	0.0302	0.0557	0.0834	
NOR	0.1125	0.0592	0.0610	0.0737	0.0649	0.0847	0.0655	0.0904	0.0853	0.1319	0.0800	0.1033	0.1131	0.1019	0.0642	0.0053	0.1137	0.1080	0.0735	0.0541	0.0587	
SHA	0.0503	0.1428	0.0632	0.0759	0.1320	0.1035	0.1127	0.0798	0.1050	0.1344	0.1315	0.0877	0.1162	0.1374	0.0733	0.1250	0.0941	0.1080	0.0726	0.0602	0.1011	
SHP	0.1488	0.1327	0.0758	0.0999	0.1463	0.1102	0.1040	0.1165	0.1392	0.1676	0.0287	0.1314	0.1069	0.1442	0.0948	0.1214	0.1511	0.1391	0.1112	0.0806	0.0870	
SLW	0.0652	0.0811	0.0345	0.0027	0.1060	0.0782	0.0680	0.0152	0.0776	0.1104	0.0878	0.0396	0.0713	0.0797	0.1031	0.0781	0.0524	0.0159	0.0069	0.0257	0.0476	
SNOR	0.1205	0.0661	0.0719	0.0869	0.0702	0.0937	0.0787	0.1020	0.0968	0.1390	0.0974	0.1185	0.0106	0.1261	0.1256	0.0825	0.1265	0.1213	0.0886	0.0636	0.0666	
STA	0.1022	0.1314	0.0815	0.0593	0.1428	0.1029	0.1197	0.0633	0.1163	0.1340	0.1485	0.0643	0.1183	0.0989	0.1546	0.0632	0.1304	0.0699	0.0503	0.0726	0.1101	
THO	0.0898	0.1321	0.0678	0.0209	0.1548	0.1148	0.1171	0.0253	0.1200	0.1317	0.1375	0.0430	0.1178	0.1093	0.1476	0.0199	0.1309	0.0729	0.0216	0.0605	0.0921	
TRA	0.0648	0.0920	0.0512	0.0166	0.1246	0.0999	0.0900	0.0227	0.0880	0.1077	0.1053	0.0454	0.0839	0.0797	0.1197	0.0193	0.0969	0.0622	0.0320	0.0337	0.0603	
WPB	0.0562	0.0687	0.0145	0.0293	0.0744	0.0579	0.0545	0.0380	0.0470	0.1096	0.0661	0.0663	0.0599	0.0652	0.0858	0.0317	0.0684	0.0788	0.0675	0.0452	0.0166	
WPC	0.1046	0.0770	0.0406	0.0540	0.0870	0.0688	0.0535	0.0650	0.0765	0.1347	0.0694	0.0946	0.0657	0.1075	0.0930	0.0550	0.0728	0.1162	0.1023	0.0726	0.0243	

The negative F_{ST} value is equal to zero; the highest values are in bold and the lowest are in bold italics.

Table S2. The individual assignment success rates for each horse breed, achieved by using eight different assignment methods in GENECLASS

Code	Sample size	Criterion															
		Bayesian method				Frequency method				Distance method							
		Rannala and Mountain (1997)	Baudouin and Lebrun (2000)	Paetkau et al. (1995)	Nei's standard distance (1972)	Nei's minimum distance (1973)	Nei's distance (1983)	Cavalli-Sforza and Edwards' distance (1967)	Goldstein's et al. distance (1995)								
		Correctly assigned	Assignment success (%)	Correctly assigned	Assignment success (%)	Correctly assigned	Assignment success (%)	Correctly assigned	Assignment success (%)	Correctly assigned	Assignment success (%)	Correctly assigned	Assignment success (%)	Correctly assigned	Assignment success (%)	Correctly assigned	Assignment success (%)
MIN	24	20	83.33	20	83.33	19	79.17	19	79.17	20	83.33	18	75.00	11	45.83		
WPC	58	53	91.38	49	84.48	47	81.03	49	84.48	46	79.31	9	15.52				
WPB	87	60	68.97	50	57.47	35	40.23	35	40.23	21	24.14	7	8.05				
SHP	95	89	93.68	88	92.63	83	87.37	91	95.79	91	95.79	43	45.26				
HUC	100	98	98.00	97	97.00	92	92.00	96	96.00	96	96.00	46	46.00				
HAF	100	96	96.00	95	95.00	91	91.00	99	99.00	99	99.00	62	62.00				
CSP	74	52	70.27	44	59.46	38	51.35	38	51.35	34	45.95	13	17.57				
MOW	36	33	91.67	33	91.67	26	72.22	31	86.11	27	75.00	16	44.44				
TRA	32	25	78.13	24	75.00	22	68.75	23	71.88	20	62.50	9	28.13				
KLA	100	98	98.00	97	97.00	92	92.00	98	98.00	98	98.00	45	45.00				
SHA	100	96	96.00	95	95.00	87	87.00	94	94.00	94	94.00	47	47.00				
LIP	10	10	100.00	10	100.00	10	100.00	10	100.00	10	100.00	6	60.00				
STA	100	90	90.00	89	89.00	85	85.00	89	89.00	88	88.00	55	55.00				
KIN	100	76	76.00	71	71.00	64	64.00	68	68.00	61	61.00	10	10.00				
CZW	100	35	35.00	29	29.00	21	20.00	12	12.00	5	5.00	5	5.00				
SLW	100	36	36.00	25	25.00	20	20.00	18	18.00	13	13.00	7	7.00				
ARAB	100	90	90.00	90	90.00	85	85.00	90	90.00	89	89.00	59	59.00				
THO	100	90	90.00	87	87.00	74	73.00	94	94.00	93	93.00	40	40.00				
CMB	100	97	97.00	97	97.00	93	93.00	96	96.00	95	95.00	51	51.00				
SNOR	100	86	86.00	85	85.00	77	77.00	93	93.00	90	90.00	30	30.00				
NOR	83	63	75.90	58	69.88	49	59.04	42	50.60	32	38.55	24	28.92				
IRI	43	42	97.67	42	97.67	38	88.37	41	95.35	41	95.35	16	37.21				
Mean			83.59	80.69	80.39	73.32	73.13	77.55	73.30								

Table S3. The confusion matrix of the Bayesian method (the Rannala and Mountain approach), as calculated by GENECLASS for each Czech-registered horse breed ($n = 1\,742$; accuracy 83.6%)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	
90	0	0	1	0	0	0	0	0	0	0	0	0	6	0	1	0	0	0	1	1	0	a = ARAB
0	97	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	b = CMB
1	1	52	3	0	2	0	0	0	0	0	0	0	0	1	1	1	0	0	1	11	0	c = CSP
1	0	2	35	0	0	0	9	0	0	0	3	0	0	0	28	0	1	8	7	6	0	d = CZW
0	1	0	2	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	e = HAF
0	0	0	1	0	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	f = HUC
0	0	0	0	1	0	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	g = IRI
1	0	1	5	0	0	0	76	0	0	0	4	0	0	0	2	0	0	7	3	0	1	h = KIN
0	0	1	0	0	0	0	0	98	1	0	0	0	0	0	0	0	0	0	0	0	0	i = KLA
0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	j = LIP
0	0	0	0	0	0	0	0	0	0	20	0	0	0	3	0	0	0	0	0	1	0	k = MIN
0	0	0	0	0	0	0	1	0	0	0	33	0	0	0	1	0	0	1	0	0	0	l = MOW
0	1	2	0	0	0	1	0	0	0	0	0	63	0	0	0	13	0	0	1	1	1	m = NOR
1	0	0	1	0	0	0	0	0	0	0	0	0	96	0	0	0	0	0	1	1	0	n = SHA
0	0	1	0	0	1	0	0	0	0	2	0	0	0	89	0	0	0	0	1	1	0	o = SHP
0	0	2	35	0	1	0	8	1	0	0	6	0	0	0	36	0	0	6	5	0	0	p = SLW
0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	86	0	0	0	1	0	q = SNOR
5	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	90	0	1	0	1	r = STA
0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	2	0	0	90	4	0	0	s = THO
1	0	0	1	0	0	0	1	0	1	0	0	1	0	0	2	0	0	0	25	0	0	t = TRA
0	0	7	2	0	2	1	1	1	0	1	0	0	0	0	4	0	0	0	1	60	7	u = WPB
0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	53	v = WPC

Table S4. The confusion matrix of the Bayes Network classifier, as calculated by WEKA for each Czech-registered horse breed ($n = 1\,742$; accuracy 78.4%)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	
85	0	1	7	0	0	0	0	0	0	0	0	1	5	0	0	0	0	0	0	1	0	a = ARAB
0	93	0	1	2	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	1	b = CMB
1	0	46	15	0	3	0	0	1	0	0	0	0	1	2	2	0	0	0	0	2	1	c = CSP
1	0	2	57	0	0	0	0	0	0	0	5	0	0	0	30	0	0	5	0	0	0	d = CZW
0	1	0	1	96	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	e = HAF
0	0	0	3	0	94	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	f = HUC
0	1	0	1	1	0	38	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	g = IRI
0	0	0	36	0	0	0	60	0	0	0	0	0	0	0	2	0	0	2	0	0	0	h = KIN
0	0	0	4	0	0	0	1	94	1	0	0	0	0	0	0	0	0	0	0	0	0	i = KLA
0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	1	0	0	0	0	0	0	j = LIP
0	0	0	2	0	0	0	0	0	0	17	0	0	0	4	0	0	0	0	0	1	0	k = MIN
0	0	0	11	0	0	0	0	0	0	0	22	0	0	0	1	0	0	2	0	0	0	l = MOW
0	2	2	2	2	0	0	0	0	0	0	0	55	0	0	1	19	0	0	0	0	0	m = NOR
3	0	0	3	0	0	0	0	0	0	0	0	0	93	0	1	0	0	0	0	0	0	n = SHA
0	0	0	3	0	1	0	0	0	0	0	1	0	0	89	0	0	0	0	0	0	1	o = SHP
0	0	0	45	0	1	0	1	1	0	0	1	0	0	0	49	0	0	2	0	0	0	p = SLW
0	0	0	1	1	0	0	0	0	0	0	0	9	0	0	0	89	0	0	0	0	0	q = SNOR
5	0	0	11	0	1	0	0	0	0	0	0	0	0	0	3	0	79	0	0	1	0	r = STA
0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	10	0	0	73	0	0	0	s = THO
0	0	0	9	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	20	0	0	t = TRA
0	1	2	13	1	3	1	0	2	0	2	0	0	2	0	2	1	0	0	0	53	4	u = WPB
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	54	v = WPC

Table S5. The confusion matrix of the Bayesian clustering method, as calculated by STRUCTURE for each Czech-registered horse breed ($n = 1\,742$; accuracy 58.03%)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	
14	0	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	a = ARAB
1	35	3	1	1	1	3	1	1	1	1	0	0	1	1	1	1	0	1	1	1	2	b = CMB
3	8	31	1	3	2	10	2	2	2	3	2	1	2	3	3	2	2	1	1	1	2	c = CSP
6	1	1	75	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	d = CZW
1	1	1	1	78	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	e = HAF
1	1	1	0	1	85	1	1	1	1	0	1	0	1	1	1	1	1	2	1	1	1	f = HUC
2	2	10	2	3	2	30	1	2	1	2	1	1	1	2	2	3	1	1	1	1	1	g = IRI
0	0	0	0	0	0	0	23	1	0	1	0	1	1	1	1	1	3	0	0	0	0	h = KIN
0	0	0	0	0	0	0	1	15	0	1	1	1	2	3	3	1	3	0	0	1	0	i = KLA
1	1	1	1	0	1	1	1	1	46	1	43	0	1	1	1	1	1	0	1	1	1	j = LIP
1	1	1	1	1	1	1	1	2	1	80	1	1	1	1	1	4	1	0	1	1	1	k = MIN
0	0	1	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	l = MOW
1	2	1	1	1	1	1	2	2	1	1	1	75	1	1	1	5	2	1	1	1	1	m = NOR
1	1	2	1	1	1	2	5	6	1	1	2	1	49	7	5	2	9	1	1	1	1	n = SHA
1	1	3	1	1	1	3	6	9	1	2	2	2	9	23	16	2	12	2	1	1	2	o = SHP
1	1	2	1	1	1	3	7	7	1	1	2	2	8	16	27	2	11	1	1	1	2	p = SLW
1	1	1	1	1	0	2	1	2	1	7	1	1	1	2	1	74	1	1	1	1	1	q = SNOR
0	1	1	1	1	0	1	5	6	0	1	1	1	5	4	4	1	67	1	1	0	0	r = STA
1	2	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	77	2	3	2	s = THO
1	1	1	1	1	2	1	1	1	2	1	1	1	1	1	1	1	1	2	67	17	1	t = TRA
1	2	1	1	1	2	1	1	1	1	0	1	1	1	1	1	1	1	3	18	43	1	u = WPB
1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	33	v = WPC